

# A Fluid Approximation for Service Systems Responding to Unexpected Overloads

Ohad Perry, Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699  
{op2105, ww2040}@columbia.edu

In a recent paper we considered two networked service systems, each having its own customers and designated service pool with many agents, where all agents are able to serve the other customers, although they may do so inefficiently. Usually we want the agents to serve only their own customers, but we want an automatic control that activates serving some of the other customers when an unexpected overload occurs. Assuming that we will not know which class will experience the overload or the timing and extent of the overload, we proposed a queue-ratio control with thresholds: When a weighted difference between the queue lengths crosses a pre-specified threshold, with the weight and the threshold depending on the class to be helped, serving the other customers is activated, so that a certain queue ratio is maintained, by only observing the queue lengths. We then developed a simple deterministic fluid approximation in which this control was shown to be optimal, and we showed how to calculate the control parameters. In this sequel, we focus on the fluid approximation itself, and develop its transient equations, which depend on a heavy-traffic averaging principle. Although deterministic, the new fluid model includes stochastic balance equations. The full fluid model here enables us to analyze scenarios we could not analyze previously. We also use this new fluid approximation to refine our previous results. Simulation experiments show that the refinements greatly increases the accuracy of our performance predictions.

*Key words:* service systems; unexpected overloads; fluid approximation; averaging principle; large-scale service systems; heavy traffic; efficiency-driven regime

*History:* This version is July 10, 2009.

---

## 1. Introduction

**Responding to unexpected overloads.** In Perry and Whitt (2009a) we considered how two service systems that normally operate independently, such as call centers, can help each other when one encounters an unexpected overload and is unable to immediately increase its own staffing. We assumed that each service system has a service pool with many agents, each of whom has the ability

to serve customers from the other system as well as its own, even though the other customers may be served inefficiently. The goal was to find a way to automatically respond to overloads, without knowledge of the arrival rates, while producing only negligible sharing under normal loads.

Toward that end, we proposed a queue-ratio control with thresholds (QR-T), which activates serving customers from the other system when a weighted difference of the two queue lengths (numbers of waiting customers) exceeds a threshold, allowing sharing in only one direction at any time. There is a target queue-ratio function and threshold for each direction of sharing. The general QR-T control allows the queue ratio to be a function of the two queue lengths, but it often suffices to use fixed queue ratios (FQR-T), which is advantageous because the control then has fewer parameters, namely the two ratio parameters and the two thresholds.

These queue-ratio controls are modification of ones proposed previously in Gurvich and Whitt (2009a,b,c). The thresholds and the application to respond to unexpected overloads are new. These QR controls tend to be effective because they simplify the problem by reducing the dimension. With the QR-T control, the two system queues tend to evolve independently when the sharing is not activated (under normal loads), but the two queues tend to evolve together in a fixed relation when the sharing is activated (under overloads). Indeed, under overloads, the two system queues tend to evolve dependently to the maximum extent. This maximum dependence can be formalized by the notion of **state-space collapse** (SSC), as in Bramson (1998), Dai and Tezcan (2005), Gurvich and Whitt (2009a).

**The Markovian X model.** To analyze this QR-T overload control and determine appropriate control parameters, we considered a Markovian X call-center model, having two customer classes, each with its own queue, and two service pools, each with many agents; see Aksin et al. (2007), Gans et al. (2003), Garnett and Mandelbaum (2000) for background on the basic call-center models. To capture a common motivation for having the systems operate independently under normal loads, we assumed the *strong inefficient-sharing condition*

$$\mu_{1,1} > \mu_{1,2} \quad \text{and} \quad \mu_{2,2} > \mu_{2,1}, \quad (1)$$

where  $\mu_{i,j}$  denotes the service rate of a class- $i$  customer by a type- $j$  agent; i.e., customers tend to get served faster by their own service pool than by the other service pool. However, many of the previous results hold under the weaker *basic inefficient-sharing condition*  $\mu_{1,1}\mu_{2,2} \geq \mu_{1,2}\mu_{2,1}$ .

**A fluid model with convex costs.** In order to determine appropriate queue-ratio functions and to approximate the performance of this QR-T control, we introduced a convex-cost framework and a simple deterministic fluid approximation for the X call-center model. Within that framework, we proved that properly chosen queue-ratio functions minimize the average steady-state cost during an overload incident, without requiring knowledge of the arrival rates. Moreover, we showed how to calculate the optimal queue-ratio functions. In addition, we indicated how to determine the thresholds. We then applied simulation to show that the optimal control for the fluid model is effective for the original stochastic X call-center model. Indeed, the simulations show that the proposed queue-ratio control with thresholds outperforms the optimal fixed partition of the servers given known fixed arrival rates during the overload, even though the proposed control does not use information about the arrival rates.

**New refined approximations.** The purpose of the present paper is to develop refined approximations for the performance of the X call-center model during overload incidents. The overloading tends to put a large service system in the efficiency-driven (ED) many-server heavy-traffic (MSHT) limiting regime, so that we can accurately approximate both the transient and steady-state performance by deterministic fluid models and diffusion-process refinements appropriate for the ED MSHT regime, as in Garnett et al. (2002), Whitt (2004, 2006). (We review the MSHT ED regime in §2.) As a consequence, we produce effective fluid approximations for the transient mean values and normal approximations for the steady-state distribution of  $(Q_1(t), Q_2(t))$ , whose accuracy is confirmed by simulation experiments. We show some of these simulation results; other simulation results appear in the online e-companion and an additional online supplement, available on the authors' web pages.

As we explain in §2, and intend to show in Perry and Whitt (2009c), the approximations we develop in this paper can be based on MSHT stochastic-process limits involving the concept of a

heavy-traffic **averaging principle (AP)**, as in Hunt and Kurtz (1994), Coffman et al. (2005). However, we prove no limit theorems here. Instead, here we contribute by showing how the SSC and the AP can be applied directly as engineering principles to better understand how large systems perform, and heuristically develop useful quantitative performance approximations. We empirically justify the approximations we develop by making extensive comparisons with simulations.

Here we take the control parameters to be given, without introducing the cost framework used in Perry and Whitt (2009a). To more carefully analyze the resulting performance, we propose a more elaborate (full) fluid model that includes a description of the transient dynamics as well as the steady-state behavior. (Previously, we only considered the steady-state behavior of the fluid model, derived simply by equating the rate in to the rate out.) The transient dynamics associated with the full fluid model involve the AP. A quick explanation is that, because of the large scale (many servers and high arrival rate), some phenomena occur in a much faster time scale than other phenomena. As a consequence, we can approximate the behavior of the rapidly changing phenomena by their local steady-state averages, which greatly simplifies the analysis of the complex system. For stochastic models, the most familiar example of the AP lies in the analysis of the large-time-scale dynamics in the theory of nearly-completely-decomposable (NCD) Markov chains, as in Courtois (1977); see Remark 2.4.1 of Whitt (2002) for more discussion and references.

**The rest of the Paper.** In §2 we specify the model and review the ED MSHT limiting regime. In §3 we develop the deterministic fluid approximation for the steady-state performance. There are two important cases of unbalanced overloads: (1) when both classes are overloaded after sharing, and (2) when one class is overloaded and the other class remains underloaded even after the sharing. In §4 we develop the deterministic fluid approximation for the transient behavior in the fully-overloaded case; that is a system of ordinary differential equations (ODE's) based on the AP. In §5 we justify the approximation developed in §4 for the transient behavior of the system during an overload by comparing the predicted performance to simulation results for the transient behavior of the original queueing system using the QR-T control. In §6 we develop two stochastic refinements to the deterministic approximations for the steady-state quantities. In §7 we compare

our approximations for steady-state mean values with simulations. In §8 we briefly describe a diffusion approximation to generate normal approximations for the full steady-state distributions. Finally, in §9 we draw conclusions.

Additional material appears in an online e-companion (EC) and an online supplement maintained on the authors' web pages. First, in §EC.1 we elaborate on the diffusion approximation, and the resulting normal approximation for the steady-state queue-length vector, briefly described in §8. In §EC.2 we compare these normal distribution approximations with simulations. Next, in §EC.3 we develop an ODE based on the AP to analyze the initially normally-loaded (without an overload)  $X$  model operating under FQR without the thresholds, and show that this ODE predicts the bad performance without these modifications previously shown in §4.1 of Perry and Whitt (2009a) through simulation. In order to do so, we develop the more complex six-dimensional ODE characterizing the fully overloaded system with two-way sharing and two thresholds.

In the online supplement we present additional simulation results, including (i) additional details about the bad performance of FQR without thresholds, (ii) the performance of FQR with one-way sharing, but no thresholds, (iii) the performance of FQR-T under normal loads, and (iv) additional simulation results to provide a more complete description of the performance of FQR-T.

## 2. Preliminaries

**The model.** We now specify the Markovian  $X$  model under consideration; it is depicted in Figure 1 of the previous paper. There are two customer classes, with customers from each class arriving according to a Poisson process. There is a queue for each customer class, from which customers are served in order of arrival. We assume that waiting customers have limited patience. A class- $i$  customer will abandon if he does not start service before a random time that is exponentially distributed with mean  $1/\theta_i$ . There are two service pools, with pool  $j$  having  $m_j$  homogeneous servers working in parallel. The service times are mutually independent exponential random variables, but the mean may depend on both the customer class and the service pool. The mean service time for a class- $i$  customer served by a type- $j$  agent is  $1/\mu_{i,j}$ . As usual, the service times, abandonment

times and arrival processes are mutually independent. Let  $Q_i(t)$  be the number of class- $i$  customers in queue and let  $Z_{i,j}(t)$  be the number of type- $j$  agents busy serving class- $i$  customers, at time  $t$ . With the assumptions above, the stochastic process  $\{(Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2) : t \geq 0\}$  is a six-dimensional continuous-time Markov chain, given any routing policy that depends on the six-dimensional state.

We are using this model to describe the system during the overload incident. We assume that customers from the two classes arrive during the overload with constant arrival rates  $\lambda_1$  and  $\lambda_2$ , which make at least one class overloaded. Our goal is to develop approximations for the stochastic process  $\{(Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2) : t \geq 0\}$  during the overload incident.

In Perry and Whitt (2009a,b) we make a strong case for focusing on the steady-state random quantities  $Q_i \equiv Q_i(\infty)$  and  $Z_{i,j} \equiv Z_{i,j}(\infty)$  in the presence of overloads. First, the abandonment always keeps the system stable, so steady state always exists. Second, we showed that steady state tends to be approached relatively quickly. Thus, we are especially interested in the random vector  $(Q_1, Q_2)$ . We show that  $(Q_1, Q_2)$  can be approximated by a bivariate normal distribution, having correlation 0 without unbalanced overloads (when there is no sharing) and having correlation 1 under unbalanced overloads (when there is sharing). We develop explicit formulas for the means and variances.

**The FQR-T control.** The FQR-T control is based on two nonnegative thresholds  $\kappa_{1,2}$  and  $\kappa_{2,1}$  and two positive queue-ratio parameters  $r_{1,2}$  and  $r_{2,1}$ . We define two (weighted) queue-difference stochastic processes  $D_{1,2}(t) \equiv Q_1(t) - r_{1,2}Q_2(t)$  and  $D_{2,1}(t) \equiv r_{2,1}Q_2(t) - Q_1(t)$ . As long as  $D_{1,2}(t) \leq \kappa_{1,2}$  and  $D_{2,1}(t) \leq \kappa_{2,1}$ , we do not allow any sharing, i.e., we only let agents serve customers from their designated class. (Note that ordinary FQR without thresholds corresponds to  $\kappa_{1,2} = \kappa_{2,1} = 0$ .)

However, pool-2 agents are allowed to start serving class-1 customers when  $D_{1,2}(t) > \kappa_{1,2}$ , provided that no pool-1 agents are still serving a class-2 customer. (We restrict attention to one-way sharing, i.e., sharing in only one direction at a time, but either direction is possible.) Pool 2 is allowed to begin service as soon as no pool-1 agents are serving class-2 customers and  $D_{1,2}(t) > \kappa_{1,2}$ . As soon as the first pool-2 agent is assigned to serve a class-1 customer, we drop the threshold

$\kappa_{1,2}$ , but keep the other threshold  $\kappa_{2,1}$ . Thus, once one-way sharing has been activated with pool 2 helping class 1, we use ordinary FQR with ratio parameter  $r_{1,2}$ : Upon service completion, a newly available type-2 agent serves the customer at the head of the class-1 queue (the class-1 customer who has waited the longest) if  $D_{1,2}(t) > 0$ ; otherwise the agent serves a customer from his own class. Only one-way sharing in this direction will be allowed until either the class-1 queue becomes empty or the other difference process crosses the other threshold, i.e., when  $D_{2,1}(t) > \kappa_{2,1}$ . As soon as either of these events occurs, newly available pool-2 agents are only assigned to class 2 and the threshold  $\kappa_{1,2}$  is reinstated. And similarly in the other direction.

Even though we intend to drop the threshold  $\kappa_{1,2}$  when sharing is activated with pool 2 helping class 1 (in the manner just described), we consider a centering constant after sharing, which can be interpreted as a threshold  $\kappa_{1,2}$ . Perry and Whitt (2009a) show that in some cases it is actually optimal to use the *shifted FQR-T control*, i.e., keeping the queues at a fixed ratio centered about a constant. Such is the case, for example, when the holding cost is separable and quadratic, i.e., of the form  $C(Q_1, Q_2) = C_1(Q_1) + C_2(Q_2)$ , where  $C_i(Q_i) = a_i + b_i Q_i + c_i Q_i^2$ ; see §EC.4 in Perry and Whitt (2009a). In these cases the optimal relation between the queues is  $Q_1 + r_{1,2} Q_2 = \kappa_{1,2}$  or  $Q_1 + r_{2,1} Q_2 = \kappa_{2,1}$  for some  $\kappa_{1,2}, \kappa_{2,1} \in \mathbb{R}$ , depending on the direction of sharing; explicit formulas appear in EC.11 and EC.12 of Perry and Whitt (2009a). If  $b_i = 0$  for  $i = 1, 2$ , then the two centering constants take the form  $\kappa_{1,2} = \kappa_{2,1} = 0$ , and we have ordinary FQR once sharing has been activated in some direction.

Here we aim to more carefully describe the performance of the system in a single overload incident. The class that is more overloaded will typically change in successive overload incidents. Without loss of generality, we assume that class-1 is the more overloaded class in the overload incident we are considering, so that we focus attention on the single queue-difference stochastic process  $D_{1,2}(t)$  and drop the subscripts on  $D$ ,  $r$  and  $\kappa$ .

**The ED MSHT limiting regime.** We intend to show that the fluid approximations developed here are asymptotically correct in the ED MSHT limiting regime in Perry and Whitt (2009c),

but we do not establish any limits here. Nevertheless, this ED MSHT limiting regime helps to understand how we get the approximations and when they should perform well.

The MSHT regimes are specified by considering a sequence of models indexed by  $n$ ; we let a superscript denote the quantity associated with model  $n$ . The main idea is that the system scale should grow. Accordingly, we assume that the arrival rates and number of servers grow proportionally to  $n$ :

$$\frac{\lambda_i^{(n)}}{n} \rightarrow \bar{\lambda}_i \quad \text{and} \quad \frac{m_j^{(n)}}{n} \rightarrow \bar{m}_j \quad \text{as} \quad n \rightarrow \infty, \quad (2)$$

where  $\bar{\lambda}_i$  and  $\bar{m}_j$  are positive constants for  $i = 1, 2$  and  $j = 1, 2$ . The individual abandonment-rates  $\theta_i$ , and service-rates  $\mu_{i,j}$  remain constant for all  $n$ .

For a Markovian I model, having one service pool, one customer class and customer abandonment, i.e., the  $M/M/m + M$  model, three different MSHT limiting regimes were identified in Garnett et al. (2002): If the system is asymptotically overloaded, then it is called the *efficiency-driven* (ED) limiting regime; if the system is asymptotically critically loaded, then it is called the *quality-and-efficiency-driven* (QED) limiting regime; if the system is asymptotically underloaded, then it is called the *quality-driven* (QD) limiting regime. These same cases without abandonment had been specified by Halfin and Whitt (1981). For one class and one pool, it is natural to let  $n$  be the total number of servers ( $m_n = n$  for all  $n$ , so that  $\bar{m} = 1$  in (2)). Then the regimes are determined by the limit  $(1 - \rho^{(n)})\sqrt{n} \rightarrow \beta$  as  $n \rightarrow \infty$ , where  $\rho^{(n)} \equiv \lambda^{(n)}/n\mu$  is the traffic intensity in model  $n$ . The regimes (i) ED, (ii) QED, and (iii) QD then occur, respectively, if the limit holds with (i)  $\beta = -\infty$ , (ii)  $-\infty < \beta < \infty$ , and (iii)  $\beta = +\infty$ .

We will be concentrating on overloaded systems, i.e., the ED regime, which for the  $I$  model is discussed in Whitt (2004). That provides important background for our work on the  $X$  model here. In that context we will consider the ED regime under the analog of the conventional more restrictive condition that  $\rho^{(n)} = \rho > 1$ . With customer abandonment, the ED regime is quite practical because the queue lengths have proper steady-state distributions whenever the abandonment rates are positive.

Before the overload occurs, a well-managed large service system will be normally loaded, which usually means that the system will be in either the QD or QED MSHT regime, but then the new arrival rates associated with the overload shift the system into the ED regime.

**Fluid limits for scaled processes.** We now indicate the kind of stochastic-process limits that should hold as  $n \rightarrow \infty$  in the ED MSHT limiting regime specified by (2) with  $\rho_i^{(n)} = \rho_i > 1$  for at least one class  $i$  (making the system overloaded for at least one class). Our descriptions will be useful when the system remains overloaded for at least one class after the sharing.

First, for deterministic fluid limits, we consider the scaled processes

$$\bar{Q}_i^{(n)}(t) \equiv \frac{Q_i^{(n)}(t)}{n} \quad \text{and} \quad \bar{Z}_{i,j}^{(n)}(t) \equiv \frac{Z_{i,j}^{(n)}(t)}{n}, \quad t \geq 0. \quad (3)$$

We anticipate that these scaled processes converge as  $n \rightarrow \infty$ , with

$$(\bar{Q}_i^{(n)}(t), \bar{Z}_{i,j}^{(n)}(t), i = 1, 2; j = 1, 2) \Rightarrow (\bar{Q}_i(t), \bar{Z}_{i,j}(t), i = 1, 2; j = 1, 2) \quad \text{as} \quad n \rightarrow \infty, \quad (4)$$

where  $\Rightarrow$  denotes convergence in distribution and the limit  $(\bar{Q}_i(t), \bar{Z}_{i,j}(t), i = 1, 2; j = 1, 2)$  evolves as a deterministic dynamical system, in particular, as a six-dimensional *ordinary differential equation* (ODE) or system of ODE's. We emphasize that the overloaded ED regime is essential for this limit to be meaningful. In the QD and QED regimes (with appropriate initial conditions) we expect the fluid limit to be  $\bar{Q}_i(t) = 0$ . With FQR-T, we should have minimal sharing when the system is not overloaded, hence we also should have  $\bar{Z}_{i,j}(t) = 0$  for all  $i, j$  with  $i \neq j$  and  $t \geq 0$ .

The limit in (4) is referred to as a weak law of large numbers (WLLN) or a functional WLLN (FWLLN). From this asymptotic perspective, we think of our deterministic fluid approximation as being  $Q_i^{(n)}(t) \approx n\bar{Q}_i(t)$  and  $Z_{i,j}^{(n)}(t) \approx n\bar{Z}_{i,j}(t)$ .

Moreover, we anticipate that all processes have well-defined steady-state limits as  $t \rightarrow \infty$  and that the double limit in (4) as  $n \rightarrow \infty$  and  $t \rightarrow \infty$  (in any order) is valid and equals the limit as  $t \rightarrow \infty$  of the ODE, which is the unique stationary point for the ODE (but none of that will be proved here). From this asymptotic perspective, we think of our deterministic fluid approximation for the steady-state random variables  $Q_i$  and  $Z_{i,j}$  as being  $Q_i \equiv Q_i^{(n)}(\infty) \approx n\bar{Q}_i(\infty)$  and  $Z_{i,j} \equiv Z_{i,j}^{(n)}(\infty) \approx$

$n\bar{Z}_{i,j}(\infty)$  for all suitably large  $n$ , but we will not include the  $n$  in our heuristic development of the approximations.

**Refined stochastic limits.** We anticipate that there also will be associated stochastic limits that serve as refinements of the fluid limits above. For these, we introduce the new scaled processes

$$\hat{Q}_i^{(n)}(t) \equiv \frac{Q_i^{(n)}(t) - n\bar{Q}_i(t)}{\sqrt{n}} \quad \text{and} \quad \hat{Z}_{i,j}^{(n)}(t) \equiv \frac{Z_{i,j}^{(n)}(t) - n\bar{Z}_{i,j}(t)}{\sqrt{n}}, \quad t \geq 0. \quad (5)$$

We anticipate that these scaled processes also converge as  $n \rightarrow \infty$ , with

$$(\hat{Q}_i^{(n)}(t), \hat{Z}_{i,j}^{(n)}(t), i = 1, 2; j = 1, 2) \Rightarrow (\hat{Q}_i(t), \hat{Z}_{i,j}(t), i = 1, 2; j = 1, 2) \quad \text{as} \quad n \rightarrow \infty, \quad (6)$$

where the limit  $(\hat{Q}_i(t), \hat{Z}_{i,j}(t), i = 1, 2; j = 1, 2)$  evolves as a stochastic process. For fully overloaded systems, this limit will be a mean-zero Gaussian process, so that the distribution is normal, specified by its covariance matrix. The limit in (6) is referred to as a functional central limit theorem (FCLT). From this new asymptotic perspective, we think of our stochastic refinement of the fluid approximation as being  $Q_i^{(n)}(t) \approx n\bar{Q}_i(t) + \sqrt{n}\hat{Q}_i(t)$  and  $Z_{i,j}^{(n)}(t) \approx n\bar{Z}_{i,j}(t) + \sqrt{n}\hat{Z}_{i,j}(t)$ . Moreover, we again anticipate that all processes have well-defined steady-state limits as  $t \rightarrow \infty$  and that the double limit in (6) as  $n \rightarrow \infty$  and  $t \rightarrow \infty$  (in any order) is valid and equals the limit as  $t \rightarrow \infty$  of the stochastic limit in (6), which is the unique stationary distribution for the limiting stochastic process (but again none of that will be proved here). From this asymptotic perspective, we think of our refined stochastic approximation for the steady-state quantities as being  $Q_i \equiv Q_i^{(n)}(\infty) \approx n\bar{Q}_i(\infty) + \sqrt{n}\hat{Q}_i(\infty)$  and  $Z_{i,j} \equiv Z_{i,j}^{(n)}(\infty) \approx n\bar{Z}_{i,j}(\infty) + \sqrt{n}\hat{Z}_{i,j}(\infty)$  for some suitably large  $n$ .

The paper Perry and Whitt (2009c) is devoted to proving the MSHT stochastic-process limits described above, while the paper Perry and Whitt (2009b) is devoted to establishing key mathematical properties of the limiting fluid ODE  $\{(\bar{Q}_i(t), \bar{Z}_{i,j}(t), i = 1, 2; j = 1, 2) : t \geq 0\}$ , including the existence of a unique solution that converges exponentially fast to a unique stationary point as  $t \rightarrow \infty$ . Even though we do not do any proofs here, we do verify these properties empirically with simulation. We demonstrate the convergence as  $n \rightarrow \infty$  in the MSHT limit by showing the performance of the scaled processes for several values of  $n$ , in particular, for  $n = 25, 100$  and  $400$ . We see

remarkable accuracy for  $n = 400$  and surprisingly good rough approximations even for  $n = 25$ . We also see the rapid convergence to steady state as  $t \rightarrow \infty$ .

We avoid proofs partly by necessity, because the full story is too long, but also by design: We want to show that these important mathematical ideas – MSHT scaling, SSC and the AP – can be applied directly to establish important engineering results, without providing full mathematical justification. For example, the scaling itself provides very important insights. We use these insights in our fluid analysis of the system, and in choosing the thresholds for sharing.

**Two time scales.** There are two points of view associated with the spatial scaling by  $n$  in (3). The first view is that of an **inside observer** - a customer or an agent. The experience of a customer or an agent is largely independent of  $n$ , because the service-time and patience distributions are independent of  $n$ . The customer waiting times tend to be independent of  $n$  as well. Since the system is overloaded, they converge to positive limits as  $n \rightarrow \infty$  (unscaled), just as in the  $M/M/n + M$  model in the ED regime; see Whitt (2004). The waiting times initially depend on  $t$  because we are initially shifting to the new steady state associated with the overload, but they converge to positive deterministic steady-state values as  $t \rightarrow \infty$ .

On the other hand, an **outside observer** has a very different perspective, which changes with  $n$ . As the system becomes larger, the time between successive arrivals becomes shorter, as does the time between successive customer abandonments and service completions. Hence, for an outside observer, a larger system means a *faster* system.

Thus, in the same system, there exist different time scales or “clocks:” One clock measures time in units of mean service times and so is of order  $O(1)$ , while another clock is of order  $O(1/n)$ , i.e., as the system gets larger, this clock is running faster. When  $n$  is large, the faster processes will tend to rapidly approach a local steady-state (in time of order  $o(1)$ ). We will exploit the different time scales when we construct the fluid approximation.

**Scaling of the thresholds.** As discussed in §6 of Perry and Whitt (2009a), the scaling also provides insight into the choice of the thresholds to use to initiate sharing. (We are now not discussing the new centering constants associated with shifted-FQR.) On the one hand, we want

the thresholds large enough that we only rarely initiate sharing inappropriately in response to stochastic fluctuations under normal loading; i.e., we want the thresholds to be large enough to prevent sharing when sharing is not needed. On the other hand, we want the thresholds to be small enough so that we actually succeed in detecting a genuine overload. There could be a problem, because abandonment alone prevents the queues from reaching high levels. Thus, if the initial thresholds are too large, then they may never be crossed, even in the presence of overloads.

The scaling in the stochastic-process limits provide guidance for the choice of the initial thresholds. The scaling indicates that we could let the initial thresholds become smaller, relatively, as  $n$  increases, so that they do not have an effect (asymptotically) on the optimal queue ratio; i.e., these thresholds could increase with  $n$ , but we should let  $\kappa_{1,2}^{(n)}/n \rightarrow 0$ . At the same time, we want these thresholds to be large compared to the stochastic fluctuations. Since the random fluctuations should be asymptotically of order  $O(\sqrt{n})$ , we can simultaneously achieve both objectives asymptotically, if we let the initial thresholds  $\kappa_{1,2}^{(n)}$  grow like  $n^p$  for  $1/2 < p < 1$ . Then, in the fluid scale, the initial thresholds are asymptotically negligible, but at the same time these thresholds will be large compared to the  $O(\sqrt{n})$  stochastic fluctuations. This asymptotic analysis does not tell us what the initial thresholds should be in any instance, but it suggests that we should be able to find effective thresholds for large systems. We use simulation to choose appropriate thresholds and verify their effectiveness.

The thresholds we do consider during the overload incident correspond to shifted-FQR. Clearly, these thresholds should be of order  $O(n)$ , so that they appear in the fluid equations. Having these final thresholds be of order  $O(n)$  is also revealing in simulations, since it allows us to see the statistical regularity, i.e., the FWLLN. We will see that the scaled performance measures then tend to be independent of  $n$  when  $n$  is not too small.

### 3. A Simple Fluid Approximation for the Steady State of the X Model

In §5 of Perry and Whitt (2009a) we developed a simple deterministic fluid approximation for the steady-state  $(Q_i, Z_{i,j}; i = 1, 2; j = 1, 2) \equiv (Q_i(\infty), Z_{i,j}(\infty); i = 1, 2; j = 1, 2)$  of the stochastic process

$\{(Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2) : t \geq 0\}$  under the assumption that the thresholds are dropped after they have been exceeded. A minor modification of the same argument yields corresponding approximations when the new thresholds are included after sharing has been initiated (to represent shifted-FQR). We give a brief account here. As before, we do not directly consider the MSHT limiting regime specified in (2), and we do not introduce the scale-factor  $n$ , but we are thinking of that regime with a suitably large  $n$ .

In advance, we do not know which class will experience the overload and need help from the other service pool. Indeed, the direction of sharing may switch in successive overload incidents. However, without loss of generality, when we consider the behavior of the system in one particular overload incident, under an unbalanced overload, **we assume that class 1 is overloaded, and more so than class 2 if class 2 is also overloaded.** (The precise meaning of “unbalanced overload” and “more so than class 2” is specified in §3.1 below.) Hence, we need only consider the queue-difference process  $D_{1,2}(t)$ , now denoted by  $D(t) \equiv Q_1(t) - rQ_2(t)$ .

**The overloaded conditions for class 1.** When we say that class 1 is overloaded, we mean that  $\lambda_1 > m_1\mu_{1,1}$ , which is equivalent to  $\rho_1 \equiv \lambda_1/(m_1\mu_{1,1}) > 1$ , where  $\rho_1$  is the class-1 traffic intensity in isolation. In other words, we assume that class 1 with pool 1 alone would produce an  $M/M/m + M$  model in the ED regime. Since we have customer abandonment, the system is stable. From §5.1 of Perry and Whitt (2009a) or Whitt (2004), we obtain the deterministic fluid approximation for the steady-state queue length of class 1 alone, namely,  $Q_1^{alone} \approx (\lambda_1 - m_1\mu_{1,1})/\theta_1$ .

There are two cases for the less-loaded class 2 after sharing: We may either have class 2 also overloaded, but less so than class 1, or class 2 underloaded.

### 3.1. The Fully-Overloaded Case: Class 2 Overloaded After Sharing

We now describe the conditions for class 2 to be overloaded after sharing, which makes both classes overloaded. **First, class 2 might be overloaded alone.** Paralleling the analysis above, that occurs if  $\lambda_2 > m_2\mu_{2,2}$  or, equivalently, if  $\rho_2 \equiv \lambda_2/m_2\mu_{2,2} > 1$ . In that event, the steady-state queue

length of class 2 alone would be  $Q_2^{alone} \approx (\lambda_2 - m_2\mu_{2,2})/\theta_2$ . In order to have sharing (class 2 helping class 1) in steady-state, we need  $Q_1^{alone} > rQ_2^{alone} + \kappa$ .

Then, for a deterministic fluid model in steady state, there should be approximately a fixed level of sharing, with  $Z_{1,2}$  type-2 agents serving class-1 customers. Thus the queue lengths should be

$$Q_1 = \frac{\lambda_1 - (m_1\mu_{1,1} + Z_{1,2}\mu_{1,2})}{\theta_1} \quad \text{and} \quad Q_2 = \frac{\lambda_2 - (m_2 - Z_{1,2})\mu_{2,2}}{\theta_2}. \quad (7)$$

We now invoke SSC to conclude that we also should have  $Q_1 = rQ_2 + \kappa$ . (We obtain equality in the fluid model, but only an approximation for the stochastic system.) Then it is easy to see that the desired amount of sharing is the unique solution of the following linear equation in the single variable  $Z_{1,2}$ :

$$Q_1 = Q_1^{alone} - \frac{Z_{1,2}\mu_{1,2}}{\theta_1} = rQ_2 + \kappa = r \left( Q_2^{alone} + \frac{Z_{1,2}\mu_{2,2}}{\theta_2} \right) + \kappa. \quad (8)$$

Clearly, there is one and only one value of  $Z_{1,2}$  yielding equality, with  $D \equiv Q_1 - rQ_2 = \kappa$ , because at  $Z_{1,2} = 0$  the left side is greater than the right side, by assumption, and the left (right) side is decreasing (increasing) in  $Z_{1,2}$ , so that there must be equality for one and only one value of  $Z_{1,2}$ . If the solution yields  $Z_{1,2} > m_2$ , then the required sharing is not possible. In that event, even if all class 2 agents work on class-1 customers, the overloads can not be balanced in the desired way. However, if  $0 \leq Z_{1,2} \leq m_2$ , then we have found our desired answer. All three variables  $Q_1$ ,  $Q_2$  and  $Z_{1,2}$  can equivalently be found by solving the following **two equations** in two unknowns ( $Q_1$  and  $Z_{1,2}$ ):

$$Q_1 = \frac{\lambda_1 - (m_1\mu_{1,1} + Z_{1,2}\mu_{1,2})}{\theta_1} \quad \text{and} \quad Q_2 = \frac{Q_1 - \kappa}{r} = \frac{\lambda_2 - (m_2 - Z_{1,2})\mu_{2,2}}{\theta_2}. \quad (9)$$

The fluid equations in Perry and Whitt (2009a) arise when we set  $\kappa = 0$  in (9).

**Now suppose that class 2 alone is underloaded.** We now consider the common case in which class 2 is initially underloaded, but becomes overloaded when it helps class 1. It is easy to see that this case is also covered by the pair of equations in (9). The equations for  $Q_i$  can be interpreted as balancing the rate in with the rate out, assuming a fixed positive queue length for that class. The only requirement of the solution to (9) is that  $0 \leq Z_{1,2} \leq m_2$  and  $Q_2 \geq 0$ . We will then necessarily have  $Q_1 = rQ_2 + \kappa$ .

EXAMPLE 1. (**canonical example**) A canonical (symmetric) example for FQR-T has  $\lambda_i = 90$ ,  $m_i = 100$ ,  $\theta_i = 0.2$ ,  $\mu_{i,j} = 1$  for all  $i$  and  $j$ ,  $r = 1$  and  $\kappa_{1,2} = \kappa_{2,1} = 10$  before the overload incident, but then a shift to  $\lambda_1 = 130$  under an unexpected overload for class 1. The simple fluid approximations above yield  $Q_1 \approx 150$  and  $Q_2 \approx 0$  if we assume that there is no sharing after the overload, and then  $Q_1 \approx Q_2 \approx 50$  using simple FQR (no shift) after sharing. The associated approximate potential waiting times (expressed in mean service times) for class 1 are reduced from 1.5 to 0.5 at the expense of increasing class-2 waiting times from 0 to 0.5. (The convexity in the underlying cost function shows that this is regarded as beneficial.) Simulation shows that this is indeed what happens, approximately.

### 3.2. The Spare-Capacity Case: Class 2 Underloaded After Sharing

The remaining case is the fortunate case when the class-2 load is low when the class-1 load is unexpectedly high. Then pool 2 might be able to help class 1 without penalty. Clearly, the FQR-T control is very desirable in this case.

From the fluid model perspective (ignoring stochastic fluctuations), this case occurs if and only if we can simultaneously have  $Q_1 = \kappa$  and  $Q_2 = 0$ . Assuming that there always are available agents in pool 2, a type-2 agent immediately serves a class-1 customer, whenever an arriving customer finds  $\kappa$  customers in queue 1. Hence, we must have  $Q_1(t) \leq \kappa$ . In the fluid model, we achieve that value  $\kappa$  for  $Q_1$  if and only if  $Z_{1,2}$  serves to balance the rate in and rate out at queue 1. Since the rate into queue 1 is  $\lambda_1$ , while the rate out is  $m_1\mu_{1,1} + \kappa\theta_1 + Z_{1,2}\mu_{1,2}$ , we obtain

$$Z_{1,2} = \frac{\lambda_1 - m_1\mu_{1,1} - \kappa\theta_1}{\mu_{1,2}}, \quad (10)$$

while still allowing queue 2 to be empty; i.e., so that the rate into queue 2 is  $\lambda_2$ , which is less than or equal to the maximum rate out of queue 2, which is  $\mu_{2,2}(m_2 - Z_{1,2})$ . Consequently, we still have  $Q_2 = 0$  along with  $Q_1 = \kappa$ . We necessarily have  $Z_{1,2} < m_2$ .

## 4. The Fluid-Model System of ODE's in the Fully Overloaded Case

We will now introduce the deterministic fluid-model ODE's to approximate the transient behavior of the CTMC  $\{(Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2) : t \geq 0\}$  in the fully overloaded case considered in §3.1. From an asymptotic perspective, we think of this approximation as stemming from the FWLLN (4), but we develop the approximation directly, without considering a sequence of models. Some readers may wish to consult Perry and Whitt (2009b) for a more formal development of the transient behavior of the fluid model based on the asymptotic approach; the proof of the MSHT limit will be given in Perry and Whitt (2009c).

We start when the overload begins, at the instant the arrival rates change. The ODE's should apply to all possible initial conditions, but the standard case is for the system to be initially in steady state with the two service pools operating independently at normal levels. The sudden shift in the arrival rates causes the system to go through two transient periods. In the first transient period, the two systems continue to operate independently, with each responding to its own new arrival rate. In the fully overloaded case being considered, the first transient period ends and the second transient period begins after  $D(t)$  exceeds its threshold and sharing is initiated, with all servers in both pools busy. That is when the AP begins to operate. We will focus on this second period. The system evolves in this second transient period approaching the steady state associated with the overload. It is this new steady state, in overload, that we are primarily trying to exploit and describe. In this section we are focusing on the second transient period. (See §8 of Perry and Whitt (2009b) for discussion of the first transient period.)

### 4.1. A Heavy-Traffic Averaging Principle

The fluid model (system of ODE's) for the second transient period (and the later stochastic refinements of the steady-state approximation) depend on an AP. Here we explain it and exploit it, but we do not prove it. In §3.1 we exploited SSC to deduce the steady-state relation  $Q_1 = rQ_2 + \kappa$  in the fully-overloaded case (using shifted FQR with centering constant  $\kappa$ ), with pool 2 serving some of the class-1 customers. However, it is evident that SSC does not actually occur in such a simple

way. Instead, the queue-difference process  $D(t) \equiv Q_1(t) - rQ_2(t)$  oscillates around the centering constant  $\kappa$ . The key observation is that the queue-difference process  $D(t)$  moves back and forth across the boundary  $\kappa$  relatively quickly, because it has a strong drift pointing toward  $\kappa$  on both sides (under typical overload conditions). These boundary crossings occur in a faster time scale than the relative changes in the other processes under consideration. All processes move due to the same arrivals and service completions (which are happening quickly because of the scale), but the processes  $Q_i(t)$  and  $Z_{i,j}(t)$  tend to be large, of the same order as the number of servers, which we assume is itself large, e.g., 100. In a very small amount of time, the fluid processes  $Q_i(t)$  and  $Z_{i,j}(t)$  do not change much relative to their values, while  $D(t)$  moves rapidly between the two regions  $(-\infty, \kappa)$  and  $[\kappa, \infty)$ , and hence approximately reaches a time-dependent steady-state very rapidly.

Since the process  $D(t)$  moves back and forth across its boundary  $\kappa$  rapidly, we conclude that  $D(t)$  approximately reaches a time-dependent steady state instantaneously at each time  $t$ , where that steady-state distribution depends on the time-dependent quantities  $Q_i(t)$  and  $Z_{i,j}(t)$ . Let  $D_t(\infty)$  denote a random variable with that time-dependent steady-state distribution. We will then exploit the time-dependent probabilities  $\pi_{1,2}(X(t)) \equiv P(D_t(\infty) \geq \kappa)$ . To obtain the steady-state quantity  $D_t(\infty)$ , we introduce a new stochastic process  $D_t \equiv \{D_t(s) : s \geq 0\}$ , which is the process  $\{D(t+s) : s \geq 0\}$ , initialized at  $D(t)$ , but with the transition rates of the stochastic process  $D$ , under the extra condition that  $(Q_1, Q_2, Z_{1,2})$  remain fixed at their values at time  $t$ .

This AP allows us to regard  $D_t$  approximately as a pure-jump continuous-time Markov process (MP), with state space  $\{k + rj : k \in \mathbb{Z}, j \in \mathbb{Z}\}$ . with transition rates that depend only on the fluid-model state at time  $t$ . There are four possible transitions in each state:  $\pm 1$  and  $\pm r$ . We obtain simplification without practical sacrifice by assuming that  $r$  is rational. For rational  $r \equiv j/k$ , this is a continuous-time Markov chain (CTMC) on the state space  $\{j/k : j \in \mathbb{Z}\}$ . We multiply by  $k$  to make all the states integers. Moreover, then the CTMC can be represented as a homogeneous quasi-birth-and-death (QBD) process, as in Definition 1.3.1 and §6.4 of Latouche and Ramaswami (1999). For each  $t$ , we can apply the logarithmic reduction algorithm in §8.7 of Latouche and Ramaswami (1999) to efficiently calculate the steady-state distribution of  $D_t$ , i.e., the distribution

of  $D_t(\infty)$ . As a consequence, we can calculate the desired probabilities  $\pi_{i,j}(X(t))$ , given any state vector  $X(t) \equiv (Q_1(t), Q_2(t), Z_{1,2}(t))$ . The specific algorithm based on the QBD representation is developed in Perry and Whitt (2009b).

We now specify the transition rates of the CTMC  $D_t$  given the time  $t$  and the state  $X(t)$ , using the integer state space. Let  $\lambda_+^{(j)}(m, X(t))$ ,  $\lambda_+^{(k)}(m, X(t))$ ,  $\mu_+^{(j)}(m, X(t))$  and  $\mu_+^{(k)}(m, X(t))$  be the transition rates of the FTSMC  $D_t$  for transitions of  $+j$ ,  $+k$ ,  $-j$  and  $-k$ , respectively, when  $D_t(s) = m > \kappa$ . Similarly, we define the transitions when  $D_t(s) = m \leq \kappa$ :  $\lambda_-^{(j)}(m, X(t))$ ,  $\lambda_-^{(k)}(m, X(t))$ ,  $\mu_-^{(j)}(m, X(t))$  and  $\mu_-^{(k)}(m, X(t))$ .

First, for  $D_t(s) = m \in (-\infty, \kappa]$ , the upward rates are

$$\lambda_-^{(k)}(m, X(t)) = \lambda_1, \quad \text{and} \quad \lambda_-^{(j)}(m, X(t)) = \mu_{1,2}Z_{1,2}(t) + \mu_{2,2}Z_{2,2}(t) + \theta_2Q_2(t), \quad (11)$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 queue, caused by a type-2 agent service completion (of either customer type) or by a class-2 customer abandonment. Similarly, the downward rates are

$$\mu_-^{(k)}(m, X(t)) = \mu_{1,1}Z_{1,1}(t) + \theta_1Q_1(t) \quad \text{and} \quad \mu_-^{(j)}(m, X(t)) = \lambda_2, \quad (12)$$

corresponding, first, to a departure from the class-1 customer queue, caused by a class-1 agent service completion or by a class-1 customer abandonment, and, second, to a class-2 arrival.

Next, for  $D_t(s) = m \in (\kappa, \infty)$ , we have upward rates

$$\lambda_+^{(k)}(m, X(t)) = \lambda_1 \quad \text{and} \quad \lambda_+^{(j)}(m, X(t)) = \theta_2Q_2(t), \quad (13)$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 customer queue caused by a class-2 customer abandonment. The downward rates are

$$\mu_+^{(k)}(m, X(t)) = \mu_{1,1}Z_{1,1}(t) + \mu_{1,2}Z_{1,2}(t) + \mu_{2,2}Z_{2,2}(t) + \theta_1Q_1(t) \quad \text{and} \quad \mu_+^{(j)}(m, X(t)) = \lambda_2, \quad (14)$$

corresponding, first, to a departure from the class-1 customer queue, caused by (i) a type-1 agent service completion, (ii) a type-2 agent service completion (of either customer type), or (iii) by a class-1 customer abandonment and, second, to a class-2 arrival.

The transition rates specified above make  $D_t$  a genuine CTMC for each time  $t$  and state  $X(t)$ . As indicated above, we have developed an efficient algorithm for computing the distribution of the steady-state quantity  $D_t(\infty)$  and thus  $\pi_{1,2}(X(t))$  in Perry and Whitt (2009b).

## 4.2. The ODE's

In the previous subsection, we observed that the rate of change of the fast-time-scale queue difference process  $D_t$  depends on (i) the state  $X(t)$  and (ii) whether or not  $D_t(s) > \kappa$ . In the same way, the rates of the CTMC  $X$  at time  $t$  depend on (i)  $X(t)$  itself and (ii) the state of  $D(t)$ . Now, for the deterministic fluid approximation of the evolution of  $X(t)$ , we let the rates (derivatives) depend on (i)  $X(t)$  itself and (ii) the steady-state probability  $\pi_{1,2}(X(t)) = P(D_t(\infty) > \kappa)$ .

First, given  $Z_{i,j}(t)$  and  $\pi_{i,j}(X(t))$ , we obtain ODE's for the two queue-length processes. Let  $\dot{x} \equiv \dot{x}(t)$  denote the derivative of  $x$  evaluated at  $t$ . We let the derivative  $\dot{Q}_1(t)$  equal the rate of increase of  $Q_1(t)$  minus its rate of decrease. The rate of increase is simply the arrival rate to customer queue 1,  $\lambda_1$ . The rate of decrease is more complicated. First, there is the rate of abandonment from queue 1, which is  $Q_1(t)\theta_1$ . Second, there is the rate of decrease from queue 1 due to service completions by servers who will next take customers from queue 1, which depends on the state of the queue-difference stochastic process. Exploiting the AP, we will not focus on the actual state of the queue-difference process, but instead focus on the average state, assuming that the queue-difference process oscillates relatively rapidly compared to the other processes. We thus assume that a proportion  $\pi_{1,2}(X(t))$  of the time that the queue-difference exceeds the threshold  $\kappa$ . That portion of the decrease rate is  $\pi_{1,2}(X(t)) (m_1\mu_{1,1} + Z_{1,2}(t)\mu_{1,2} + Z_{2,2}(t)\mu_{2,2})$ . There will be corresponding, but different, rates of decrease for the proportion of time  $1 - \pi_{1,2}(X(t))$  that the queue-difference is less than or equal to the threshold  $\kappa$ . That reasoning leads to the system of three ODE's

$$\begin{aligned}
\dot{Q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(X(t)) [Z_{1,2}(t)\mu_{1,2} + Z_{2,2}(t)\mu_{2,2}] - \theta_1 Q_1(t) \\
\dot{Q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(X(t))) [Z_{2,2}(t)\mu_{2,2} + Z_{1,2}(t)\mu_{1,2}] - \theta_2 Q_2(t) \\
\dot{Z}_{1,2}(t) &\equiv \pi_{1,2}(X(t)) Z_{2,2}(t)\mu_{2,2} - (1 - \pi_{1,2}(X(t))) Z_{1,2}(t)\mu_{1,2},
\end{aligned} \tag{15}$$

More compactly, we have a single three-dimensional ODE with the general form  $\dot{X}(t) = \Psi(X(t), t)$  for a function  $\Psi$ . In addition, our ODE is *autonomous* (or *time invariant*) because  $\Psi(X(t), t) \equiv \Psi(X(t))$ . An autonomous ODE does not depend explicitly on the time-argument  $t$ , and its behavior is invariant to shifts in the time origin. Thus we propose the autonomous ODE

$$\dot{X}(t) \equiv (\dot{Q}_1(t), \dot{Q}_2(t), \dot{Z}_{1,2}(t)) = \Psi(X(t)) \equiv \Psi(Q_1(t), Q_2(t), Z_{1,2}(t)), \quad t \geq 0, \quad (16)$$

where  $\Psi : [0, \infty)^2 \times [0, m_2] \rightarrow \mathbb{R}^3$  is displayed via (15) above. The derivatives in (15) are evident given the transition rates of the CTMC, given that we replace the CTMC by an ODE and invoke the AP. A more systematic derivation, involving the asymptotic approach, appears in Perry and Whitt (2009b,c).

**Steady state.** We now characterize the steady state of the fluid model in the fully overloaded case with one-way sharing. We can directly apply the ODE for  $Z_{1,2}(t)$  to find  $\pi_{1,2}$  by noting again that in steady state  $\dot{Z}_{1,2}(t) = 0$ . Thus,

$$\pi_{1,2} = \frac{Z_{1,2}\mu_{1,2}}{Z_{1,2}\mu_{1,2} + (m_2 - Z_{1,2})\mu_{2,2}}. \quad (17)$$

This yields (9). In §6.1 we show how to use the AP to refine the fluid equations (9).

## 5. Validating the Transient Approximation through Simulation Experiments

In Perry and Whitt (2009a) we made a case for focusing on the steady-state behavior of the FQR-T control when an overload incident occurs, because these many-server systems tend to approach steady state quickly; see especially §EC.1 there. Hence the present paper is primarily devoted to developing improved approximations for the steady-state distributions. However, the control is dynamic, so that the transient behavior is important too. Our third paper Perry and Whitt (2009b) is primarily devoted to the transient behavior of the fluid model, establishing important mathematical properties and providing an efficient algorithm. Nevertheless, we now want to provide evidence that our proposed approximation is effective, not just for the steady-state values but also for the transient behavior.

Accordingly, in this section we compare numerical results for the transient behavior of the fluid model, based on our algorithm from Perry and Whitt (2009b), to simulation estimates of the actual performance measures in the original queueing model. This will show that the transient approximations are computable and sufficiently accurate for engineering applications. We will show that the deterministic fluid model does not capture important stochastic fluctuations unless the scale is very large, but the fluid model provides remarkably accurate approximations for the mean values of the key queueing processes,  $Q_1(t)$ ,  $Q_2(t)$  and  $Z_{1,2}(t)$ , provided that the scale is not too small.

In order to demonstrate the MSHT limits in the ED regime described in §2, we report results for scaled processes, as in (3), for several values of  $n$ . We will then be confirming the FWLLN in (4) via the simulation. Our simulation examples throughout the paper will have parameters related to a **base case** that we consider here as well. It has several parameters depending on  $n$ :  $m_i \equiv m_i^{(n)} = n$ ,  $\lambda_1 \equiv \lambda_1^{(n)} = 1.3n$ ,  $\lambda_2 \equiv \lambda_2^{(n)} = 0.9n$  and  $\kappa \equiv \kappa^{(n)}$ . Here we take  $\kappa^n = 0$ , but we will later also consider a positive  $\kappa$ , specifically  $\kappa \equiv \kappa^n = 0.1n$ . The other model parameters are independent of  $n$ :  $\theta_1 = \theta_2 = 0.2$ ,  $\mu_{1,1} = \mu_{2,2} = 1.0$  and  $\mu_{1,2} = \mu_{2,1} = 0.8$ . The arrival rates are chosen to put class 1 in a focused overload, while class 2 is initially normally loaded or slightly underloaded, but becomes overloaded too after the sharing. The rest of the parameters are chosen to make a symmetric model, where serving the other class is less efficient. We use the FQR-T control with ratio parameter  $r = 0.8$ ; this makes the QBD matrices be as in (23) and (24) of Perry and Whitt (2009b), following the general structure in Section 4.2 there; the algorithm is given in §9.3 there.

We have in mind large-scale applications, e.g., with  $n \geq 50$ , but to test the limits of the approximations, we also consider smaller systems. Specifically, we consider the three cases:  $n = 10$ ,  $n = 25$  and  $n = 100$ , initialized empty. Since the processes are scaled, they all have the same fluid approximation. For each  $n$ , we ran 1000 independent replications, sampling each of the 1000 simulated sample paths every  $h \equiv 0.01$  time units over the time interval  $[0, T] = [0, 50]$ . This gives 5001 sample points for each replication.

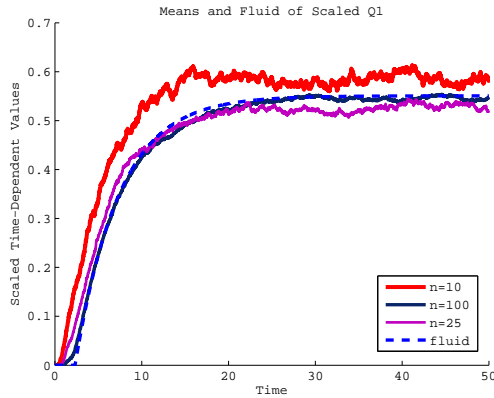
Figures 1-3 show the fluid approximation together with simulation estimates of the time-dependent mean values for each  $n$ , specifically, the averages of the 1000 observed values of three scaled processes  $\bar{Q}_i^{(n)}(t) \equiv n^{-1}Q_i^{(n)}(t)$ ,  $i = 1, 2$ , and  $\bar{Z}_{1,2}^{(n)}(t) \equiv n^{-1}Z_{1,2}^{(n)}(t)$  at each of the 5001 sample points. Figure 4 shows one sample path of  $\{\bar{Q}_1^{(n)}(t) : 0 \leq t \leq 50\}$ , when  $n = 100$ , together with the fluid approximation, to show the typical stochastic fluctuations. These stochastic fluctuations are the reason for using a large number of replications in order to accurately estimate the mean values at each point along the sample path. The statistical precision of the estimators is directly visible in the plots, because the processes are effectively in steady state in the second half of the time interval  $[0, 50]$ . As  $n$  grows larger, the impact of these fluctuations decreases; they are of order  $1/\sqrt{n}$  by (6). The stochastic fluctuations show that importance of the diffusion refinements in §8.

One immediate observation is that, consistent with the conjectured FWLLN in (4), the larger the system, the better the fluid approximates the means. The figures clearly show that  $n \geq 100$  is “large enough,” in the sense that the simulated means are extremely close to the fluid approximation. Even a relatively small system, with only 25 agents in each pool, is approximated quite well by the fluid. However, the fluid approximation is very rough when  $n = 10$ . There is approximately 25% difference between the fluid and the means of  $\bar{Q}_2^{(n)}(t)$  when  $n = 10$ .

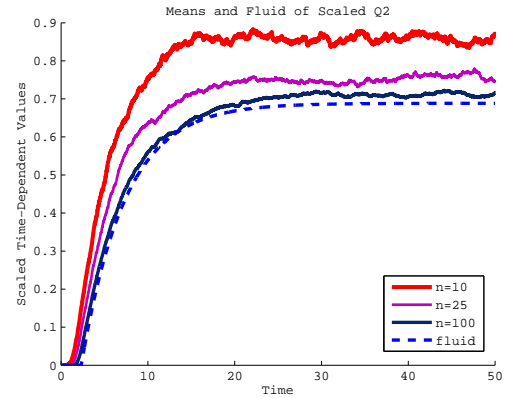
Nevertheless, the fluid approximation is useful even for small systems, because the shape of the curves of the simulation means for  $n = 10$  is the same as the shape of the fluid curve; in particular, the rate of convergence to steady state is about the same in all systems. Since the fluid approximation was shown to converge exponentially fast to steady state in §7.3 of Perry and Whitt (2009b), we see that the same must be true, approximately, for the queueing system even for quite small numbers of servers.

## 6. Stochastic Refinements to the Steady-State Fluid Approximation

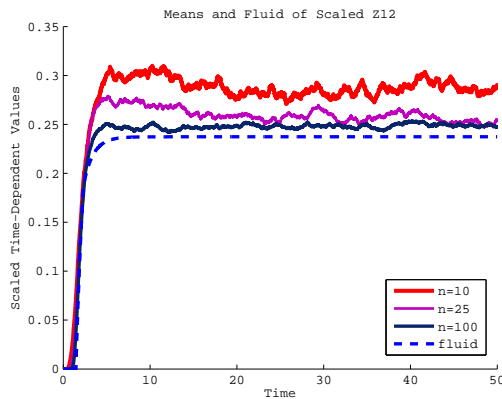
In this section we present two stochastic refinements to the deterministic fluid-model approximations for the steady-state quantities  $Q_i$  and  $Z_{1,2}$  describing performance during the overload, assuming shifted FQR is used then. The first exploits the AP to determine the average queue difference for the fully-overloaded case in §3.1. The second develops a birth-and-death-process (BD)



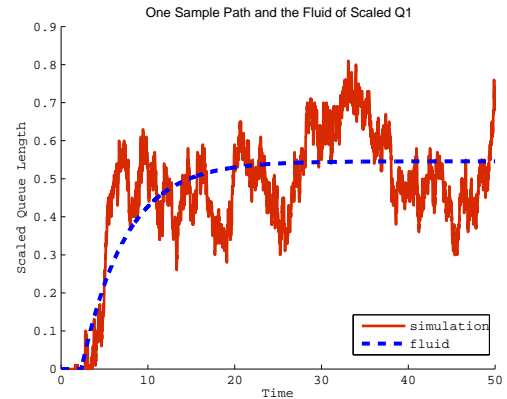
**Figure 1** A comparison of simulated estimates of  $E[\bar{Q}_1^{(n)}(t)]$  for  $n = 10, 25, 100$  to the fluid approximation in the base case.



**Figure 2** A comparison of simulated estimates of  $E[\bar{Q}_2^{(n)}(t)]$  for  $n = 10, 25, 100$  to the fluid approximation in the base case.



**Figure 3** A comparison of simulated estimates of  $E[\bar{Z}_{1,2}^{(n)}(t)]$  for  $n = 10, 25, 100$  to the fluid approximation in the base case.



**Figure 4** A comparison of one sample path of  $\bar{Q}_1^{(n)}(t)$  when  $n = 100$  to the fluid approximation in the base case.

approximation for the steady-state queue length  $Q_1$  in the spare-capacity case of §3.2.

### 6.1. The Average Difference $E[D]$ in the Fully-Overloaded Case

For the stochastic  $X$  model in the fully-overloaded case, SSC does not happen exactly; we do not get precisely  $Q_1 = rQ_2 + \kappa$ . Instead, under the overloading we are considering, the queue-difference process  $D(t)$  oscillates around the centering constant  $\kappa$ . As discussed in §4.1 above, we can apply the AP to find an approximating steady-state distribution of  $D(t)$  by treating it as a fast-time-

scale Markov process (MP). Let  $D$  denote a random variable with the limit of these steady-state distributions as  $t \rightarrow \infty$ .

We propose refining our fluid approximation for the steady-state distribution by replacing the target difference  $\kappa$  by the mean  $E[D]$ . To find  $E[D]$  we solve the balance equations of the MP process above, and then take the mean

$$E[D] = \sum_{j=-\infty}^{\infty} jP(D=j). \quad (18)$$

Since the drifts tend to point strongly toward the centering constant  $\kappa$ , it usually suffices to perform the sum for  $\kappa - 20 \leq j \leq \kappa + 20$ .

We now obtain our refined approximation, assuming that the queue difference is  $E[D]$  instead of  $\kappa$ . The calculation of  $E[D]$  can be easily done if  $Q_1$ ,  $Q_2$  and  $Z_{1,2}$  are known. Since they depend on the value  $E[D]$ , we need to solve for them simultaneously. To do that, we propose an iterative algorithm which solves the **three equations**

$$\begin{aligned} Q_1 &= \frac{\lambda_1 - (m_1\mu_{1,1} + Z_{1,2}\mu_{1,2})}{\theta_1}, & Q_2 &= \frac{Q_1 - E[D]}{r} = \frac{\lambda_2 - (m_2 - Z_{1,2})\mu_{2,2}}{\theta_2}, \\ E[D] &= \sum_{j=-k_{2,1}}^{\infty} jP(D=j). \end{aligned} \quad (19)$$

For the iterative procedure, it is natural to start with the values of  $Q_1$ ,  $Q_2$  and  $Z_{1,2}$  obtained from (9), and then calculate the distribution of  $D$  and  $E[D]$ . We can then obtain new values of  $Q_1$ ,  $Q_2$  and  $Z_{1,2}$  by solving (9) again with  $E[D]$  replacing  $\kappa$ . We then can keep iterating. Experience indicates that this iteration consistently converges in a few iterations (typically only two), yielding the solution to (19).

## 6.2. A BD-Process Refinement for the Spare-Capacity Case

For the case in which queue 2 has spare capacity, considered in §3.2, we now develop another refinement, obtaining a non-degenerate approximation for the distribution of  $Q_1$ . In this case, because of the available agents in pool 2, as soon as  $Q_1$  exceeds the centering constant  $\kappa$ , an idle pool-2 agent serves a customer from class 1. Thus, it is evident that we must have  $Q_1 \leq \kappa$ .

Because of the averaging principle, it is not hard to estimate the approximate distribution of  $Q_1$ . To do so, we observe that we can regard the class-1 queue as evolving below the level  $\kappa_{1,2}$  by itself as a BD process. When the queue length is  $j$ , the birth rate is a constant  $\lambda_1$ , while the death rate is approximately  $m_1\mu_{1,1} + \theta_1 j$ . (Queue 2 plays no role.) For the reason given, the birth rate is 0 when the queue is at  $\kappa$ . The death rate should be small when the queue length is small. For the approximation to be good, we do not want  $Q_1$  to spend much time at very low levels, like 1 or 0. That can be verified approximately by looking at the approximate BD steady-state distribution. In any case, we let the death rate be 0 when the queue length is 0. Our refined approximation for the distribution of  $Q_1$  is the steady-state distribution of this finite-state BD process.

Since  $Q_1^{alone} = (\lambda_1 - m_1\mu_{1,1})/\theta_1 > \kappa$ , the birth rate always exceeds the death rate here. Indeed, the BD process here for  $\kappa - Q_1(t)$  is stochastically bounded above by the queue-length process in an  $M/M/1/\kappa$  queue, where  $\kappa$  serves as the size of a finite waiting room. If we take the asymptotic perspective in §2, this stochastic bound shows that the difference  $\kappa - Q_1$  should be of order  $O(1)$  as  $n \rightarrow \infty$ . Hence this adjustment should be asymptotically negligible in both the diffusion scale ( $\sqrt{n}$ ) and the fluid scale ( $n$ ). However, the refinement can help in actual examples, even large ones with 1000 servers in each pool.

As a refined deterministic fluid approximation, we use the mean value of the steady-state distribution of the BD process here. However, by this method, we also obtain an estimate for the variance and the entire distribution of  $Q_1$ . The observed  $M/M/1$  structure indicates that the distribution of  $\kappa - Q_1(t)$  should be approximately a truncated geometric distribution. That is quite different from the approximate normal distribution we derive for the fully-overloaded case in §8.

## 7. Simulation Experiments to Evaluate the Steady-State Mean Values

**The overloaded case.** We have developed deterministic fluid approximations for the steady-state mean values in the fully overloaded case via the solutions to the two equations in (9) and the three equations in (19). We now compare these approximations to simulation estimates. In order to use the simulation to substantiate the conjectured stochastic-process limits in §2, we choose parameters

corresponding to scaled systems, indexed by  $n$ , letting  $n$  take the values 25, 100 and 400. We have considered much larger  $n$ , such as  $n = 1000$ , but from the results for  $n = 400$ , we see that accurate results will be obtained for all  $n$  larger than 400.

We consider the **base case**, introduced in §5, with  $r = 1$ . This makes the model symmetric and reduces the fast-scale MP to a BD process. In the EC we present corresponding results for asymmetric models. Specifically, when  $r \neq 1$  the fast-scale MP is no longer a BD process, and a more involved analysis is needed.

In all our simulation experiments, we used 5 independent runs, each with 300,000 arrivals. We report averages together with the half widths of the 95% confidence intervals, based on a  $t$  statistic with four degrees of freedom. Simulation results for the base case above are presented in Table 1 below. Table 1 shows both the steady-state mean values and the associated scaled values (i.e.,

	n=25			n=100			n=400		
perf. meas.	2 equ.	3 equ.	sim.	2 equ.	3 equ.	sim.	2 equ.	3 equ.	sim.
$E[Q_1]$	16.6	14.4	15.7 $\pm 0.3$	65.6	63.1	63.6 $\pm 1.9$	262.2	259.7	258.3 $\pm 5.0$
$E[Q_1/n]$	0.656	0.575	0.629 $\pm 0.013$	0.656	0.631	0.636 $\pm 0.019$	0.656	0.649	0.646 $\pm 0.013$
$E[Q_2]$	13.6	16.4	15.9 $\pm 0.4$	55.6	58.6	58.6 $\pm 1.8$	222.2	225.3	223.9 $\pm 5.0$
$E[Q_2/n]$	0.556	0.656	0.636 $\pm 0.016$	0.556	0.586	0.586 $\pm 0.018$	0.556	0.563	0.560 $\pm 0.013$
$E[D]$	–	–2.0	–0.2 $\pm 0.3$	–	4.6	5.0 $\pm 0.1$	–	34.4	34.4 $\pm 0.04$
$\kappa - E[D]$	–	5.0	3.2 $\pm 0.3$	–	5.4	5.0 $\pm 0.1$	–	5.6	5.6 $\pm 0.04$
$E[Z_{1,2}]$	5.3	5.8	5.6 $\pm 0.1$	21.1	21.7	21.9 $\pm 0.04$	84.4	85.1	84.2 $\pm 1.2$
$E[Z_{1,2}/n]$	0.211	0.231	0.224 $\pm 0.003$	0.211	0.217	0.219 $\pm 0.004$	0.211	0.213	0.210 $\pm 0.003$

**Table 1** A comparison of the basic fluid approximations based on two equations in (9) and its refinement based on the three equations in (19) with simulation results in the base case, having  $m_1 = m_2 = 1.0n$ ,  $\lambda_1 = 1.3n$ ,  $\lambda_2 = 0.9n$ ,  $\mu_{1,1} = \mu_{2,2} = 1.0$ ,  $\mu_{1,2} = \mu_{2,1} = 0.8$ ,  $\theta_1 = \theta_2 = 0.2$  and  $\kappa = 0.1n$  (rounding up to the nearest integer if necessary).

divided by  $n$ ). The unscaled values helps us evaluate the performance of the actual system, while the scaled values show the convergence of the stochastic-process limits in (4). Table 1 clearly shows

that the level of accuracy grows as  $n$  gets larger, but even for relatively small systems, the fluid approximation gives reasonable results.

Table 1 also gives the approximation for the steady-state mean of the unscaled weighted-difference process  $D(t)$ , as developed in §6.1, and compares it to simulation results. The sixth row in the table is especially insightful. It shows that  $E[D]$  is about the same distance from  $\kappa_{1,2}$  for each  $n$ , thus strengthening our claim that  $D(t)$  should have fluctuations of order  $O(1)$  as  $n \rightarrow \infty$ .

In closing, we remark that we rounded up the centering constant  $\kappa$  to the nearest integer when  $n = 25$ ; i.e., we used  $\kappa = 3$  when  $n = 25$ . In the table we show the fluid solution using  $\kappa = 2.5$  so as to make the scaled fluid solutions uniform. However, the solution using  $\kappa = 3$  is similar.

**Independent Cases.** One of our objectives is to avoid sharing without unbalanced overloads. That occurs in two scenarios: (i) under normal loads, and (ii) under balanced overloads. In both of these cases our FQR-T control makes the  $X$  model operate approximately as two independent  $M/M/n + M$  systems, each operating in the  $QD$  or  $QED$  regime in the first scenario (depending on the actual load of each queue), or the  $ED$  regime in the second scenario. We present supporting simulation results in the online supplement.

**The spare-capacity case.** For the spare capacity case, we modify the base case above to make queue-1 overloaded, while pool-2 has enough spare capacity to potentially serve all the extra class-1 customers. As before, we just change the arrival rates, in this case to  $\lambda_1 = 1.1n$  and  $\lambda_2 = 0.8n$ .

It is easy to see that pool 2 has spare capacity (in the fluid scale). We can analyze the available capacity from this deterministic-fluid-approximation perspective as follows: First, we observe that class 1 has an extra arrival rate of  $0.1n$ , whereas pool 2 has  $0.2n$  “extra” service rate, assuming that  $0.8n$  servers are enough to take care of all the class-2 arrivals. Since pool-2 agents serve class-1 customers at rate  $\mu_{1,2} = 0.8$ , we initially estimate that we need to have at least  $0.125n$  pool-2 agents working with class-1 customers. However, upon further analysis, we see that the number of pool-1 agents needed is actually less than that, because queue 1 will stabilize at the centering constant  $\kappa = 0.1n$ , and thus  $\theta_1 Q_1 = 0.02n$  class-1 customers will abandon. Hence, only about  $0.105n$  pool-2 agents should be needed to serve class 1. In any case, pool 2 has spare capacity.

We compare the approximation from §3.2 with simulation results in Table 2. The approximations are given in §3.2. Our initial approximation for  $Q_1$  from §3.2 is  $\kappa$ , but that is not shown in Table 2. Instead, we only show the BD refinement from §6.2. (The cruder approximation would yield values of 2.5, 10.0 and 40.0 in the first row.) We see that the refined approximation is much better for large  $n$ . For the approximation of  $Z_{1,2}$ , we use (10).

	n=25		n=100		n=400	
perf. meas.	approx.	sim.	approx.	sim.	approx.	sim.
$E[Q_1]$	1.1	3.3 $\pm 0.1$	5.2	6.4 $\pm 0.6$	29.0	30.1 $\pm 0.5$
$E[Q_1/n]$	0.04	0.13 $\pm 0.00$	0.05	0.06 $\pm 0.01$	0.07	0.07 $\pm 0.00$
$E[Q_2]$	0	3.4 $\pm 0.05$	0	2.7 $\pm 0.5$	0	1.0 $\pm 0.2$
$E[Q_2/n]$	0	0.14 $\pm 0.00$	0	0.027 $\pm 0.005$	0	0.003 $\pm 0.000$
$E[Z_{1,2}]$	2.5	3.9 $\pm 0.1$	10.0	12.2 $\pm 0.5$	40.0	43.4 $\pm 1.2$
$E[Z_{1,2}/n]$	0.100	0.156 $\pm 0.007$	0.100	0.122 $\pm 0.007$	0.100	0.108 $\pm 0.003$

**Table 2** A comparison of the approximation for the steady-state performance measures in the spare-capacity case with simulation results. The arrival rates are now  $\lambda_1 = 1.1n$  and  $\lambda_2 = 0.8n$ .

## 8. A Diffusion-Process Refinement

In the fully-overloaded case, we now go beyond the deterministic fluid approximation in §3.1 and §6.1 to obtain a diffusion-process refinement, which yields a non-degenerate approximation for the steady-state distribution of the two queue lengths. The approximating distribution is bivariate normal, where the means are the previous fluid approximations. In addition, the approximating correlation is 1 and the variances are

$$\text{Var}(Q_1) \approx \frac{r^2(\lambda_1 + \lambda_2)}{(1+r)(r\theta_1 + \theta_2)} \quad \text{and} \quad \text{Var}(Q_2) \approx \frac{(\lambda_1 + \lambda_2)}{(1+r)(r\theta_1 + \theta_2)}. \quad (20)$$

We base our approximation on a special case for which we can do the asymptotic analysis exactly, and then we extend the approximation heuristically to other cases. The special case we can analyze exactly has  $\theta_1 = \theta_2$  and  $\mu_{1,2} = \mu_{2,2}$  (with class 1 overloaded as usual). Under those additional

assumptions, the total queue length  $Q_{\Sigma}(t) \equiv Q_1(t) + Q_2(t)$  behaves the same as the queue length in the  $M/M/m+M$  model in the ED regime, as analyzed in Whitt (2004). In this special case, we can directly obtain a FCLT like (6) for the total queue-length stochastic process. From Whitt (2004), we see that the limit is an Ornstein-Uhlenbeck diffusion process with infinitesimal mean  $m(x) = -\theta_1 x$  and infinitesimal variance  $\sigma^2 \equiv \sigma^2(x) = 2(\lambda_1 + \lambda_2)$ . That diffusion process has a normal steady-state distribution. We invoke SSC to treat the individual queue lengths; that yields the correlation 1. Even though this normal-distribution approximation is not asymptotically correct in general, the simulation results are impressive; e.g., see Table EC.1 and the other tables and figures in §EC.2. For additional details, see §EC.1.

## 9. Conclusions

In this paper we further investigated the FQR-T routing policy for the  $X$  model, proposed in Perry and Whitt (2009a) as a way for two service systems to help each other during unexpected overloads. We showed how the performance of the FQR-T control can be analyzed, exploiting the fact that the overload puts the system in the ED MSHT limiting regime, reviewed in §2. Even though the approximations have a complicated basis, supported by stochastic-process limits (explained in §2, but not established here), the steady-state approximations developed in §3, §6, and §8 are relatively simple and easy to apply.

However, those approximations were not so easy to develop. From the theoretical point of view, the main contribution of this paper is the reduction of a complicated queueing model to more elementary and elegant approximate models, using the heavy-traffic averaging principle (AP) in the development of the deterministic fluid approximation (the system of ODE's in §4.1) and state-space collapse (SSC) in the diffusion approximations (and resulting approximate normal distribution in §8). The relatively simple initial fluid approximation in §3 was refined in useful ways in §6 and §8. Simulation experiments in §5, §7, the e-companion and the online supplement (on the authors' web pages) confirm that the approximations for both the transient mean values and the steady-state distributions are quite accurate.

Work is in progress to establish additional results: First, we intend to show in Perry and Whitt (2009c) that the deterministic fluid approximation is asymptotically correct in the MSHT ED regime. We also intend to establish the FCLT refinement in (6). Second, we intend to extend the performance approximations to the  $X$  model with non-exponential distributions, paralleling the previous results in Whitt (2006) for the single-class single-pool  $I$  model.

The whole discussion was limited to the two-class-two-pool  $X$ -model setting, but we believe that the control and the results can be generalized to more complex systems, as in Gurvich and Whitt (2009a,b,c). First, for applications to modern call centers, we would want the two service systems to be more general than the  $I$  models considered here. Second, we would like to consider sharing among more than two service systems. The QR-T and FQR-T controls extend quite naturally to more complex systems, but our mathematical analysis, both here and in our other papers, evidently does not extend so easily. Such extensions remain a topic for future research.

*Acknowledgments.* This research was supported by NSF Grant DMI-0457095.

## References

- Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Mgmt.* **16** 665–688.
- Bramson, M. 1998. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* 30 (1-2) 89–148.
- Coffman, E. G., A. A. Puhalskii, M. I. Reiman. 1995. Polling systems with zero switchover times: a heavy-traffic averaging principle. *Annals of Applied Probability* 5, 681–719.
- Courtois, P. 1977. *Decomposability*, Academic Press, New York.
- Dai, J. G., T. Tezcan. 2005. State space collapse in many server diffusion limits of parallel server systems. Georgia Institute of Technology, Atlanta, GA.
- Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.
- Garnett, O., A. Mandelbaum. 2000. An introduction to skill-based routing and its operational complexities. Unpublished manuscript, Technion, Haifa, Israel. <http://iew3.technion.ac.il/serveng>

- Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Opns. Mgmt.* **4** 208–227.
- Gurvich, I., W. Whitt. 2009a. Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* **34** 363–396.
- Gurvich, I., W. Whitt. 2009b. Scheduling Flexible Servers with Convex Delay Costs in Many-Server Service Systems. *Manufacturing and Service Operations Management* **11** 237–253.
- Gurvich, I., W. Whitt. 2009c. Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.*, forthcoming.  
Available at: <http://www.columbia.edu/~ww2040/recent.html>
- Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** (3), 567–588.
- Hunt, P.J., T.G. Kurtz. 1994. Large loss networks. *Stochastic Processes and their Applications* **53**, 363–378.
- Latouche, G., V. Ramsaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modelling*, SIAM and ASA, Philadelphia.
- Perry, O., W. Whitt. 2009a. Responding to unexpected overloads in large-scale service systems. *Management Sci.*, forthcoming. Available at: <http://www.columbia.edu/~ww2040/recent.html>
- Perry, O., W. Whitt. 2009b. Transient and stability analysis of the many-server heavy-traffic fluid limit for the overloaded X call-center model. Submitted to *Math. Oper. Res.*. Available at: <http://www.columbia.edu/~ww2040/recent.html>
- Perry, O., W. Whitt. 2009c. Heavy-traffic limits for the overloaded X model via an averaging principle. in preparation.
- Whitt, W. 2002. *Stochastic-Process Limits*, Springer, New York.
- Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50** (10), 1449–1461.
- Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Operations Research* **54** (1) 37–54.

**This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.**

## e-Companion

This e-companion contains material that had to be omitted from the main paper because of insufficient space. First, in §EC.1 we elaborate on the diffusion approximation to generate normal approximations for the steady-state distributions of the queue lengths, discussed in §8. In §EC.2 we compare our normal-distribution approximations with simulations. In §EC.3 we develop an ODE based on the AP to analyze the initially normally-loaded (without an overload)  $X$  model operating under FQR without the thresholds, and show that this ODE predicts the bad performance without these modifications previously shown in §4.1 of Perry and Whitt (2009a) through simulation. Additional simulation results appear in an online supplement maintained on the authors' web pages.

### EC.1. A Diffusion-Process Refinement in the Fully-Overloaded Case

In this section we elaborate on the bivariate normal approximation for the steady-state random vector of queue lengths  $(Q_1, Q_2)$  during an overload incident in the fully-overloaded case, which we presented in §8. The mean values are the steady-state fluid values in (19). The individual variances are given in (20). Because of SSC, the correlation is 1. We now elaborate.

**Leveraging known results.** We base our approximation on a special case for which we can do the asymptotic analysis exactly, and extend the approximation heuristically to other cases. The special case we can analyze exactly has  $\theta_1 = \theta_2$  and  $\mu_{1,2} = \mu_{2,2}$  (with class 1 overloaded as usual). Under those additional assumptions, the total queue length  $Q_\Sigma(t) \equiv Q_1(t) + Q_2(t)$  behaves the same as the queue length in the  $M/M/m + M$  model in the ED regime, as analyzed in Whitt (2004). Since the system is fully overloaded, as an approximation we assume that all the agents are busy all the time. (That should be asymptotically correct in the MSHT limit.) Thus, the departure rate by service completion has the constant value  $m_1\mu_{1,1} + m_2\mu_{2,2}$ . The assumption that  $\mu_{1,2} = \mu_{2,2}$  implies that it does not matter which class the type-2 agents are serving. Since the total arrival process is a superposition of two independent Poisson processes, the total arrival process is directly a Poisson process with rate  $\lambda_1 + \lambda_2$ . Finally, since  $\theta_1 = \theta_2$ , there is a common abandonment rate for both classes.

So in this special case we can directly obtain a FCLT of the form (6) for the total queue-length stochastic process, using the scaling in (5). From Whitt (2004), we see that the appropriately scaled version of the difference between the total queue length and its fluid approximation can be approximated by an Ornstein-Uhlenbeck (OU) diffusion process with infinitesimal mean  $m(x) = -\theta_1 x$  and infinitesimal variance  $\sigma^2 \equiv \sigma^2(x) = 2(\lambda_1 + \lambda_2)$ . It is well known that this OU process has a normal steady-state distribution with mean zero and variance  $Var(Q_\Sigma) \approx (\lambda_1 + \lambda_2)/\theta_1$ . At this point, we invoke the SSC to get associated limits and approximations for the individual queue lengths. (We are applying SSC without proof here.) We start from the approximation for  $(Q_1, Q_2)$  in (19). The SSC implies that the correlation is 1. Thus we invoke the SSC to get

$$Q_1 \approx \frac{r(Q_\Sigma + E[D])}{1+r} \quad \text{and} \quad Q_2 \approx \frac{Q_\Sigma - E[D]}{1+r} \quad (\text{EC.1})$$

for the steady-state variables. That gives a joint normal distribution for  $(Q_1, Q_2)$  with correlation 1 and individual variances

$$Var(Q_1) \approx \frac{r^2(\lambda_1 + \lambda_2)}{(1+r)^2\theta_1} \quad \text{and} \quad Var(Q_2) \approx \frac{(\lambda_1 + \lambda_2)}{(1+r)^2\theta_1}. \quad (\text{EC.2})$$

**A heuristic extension.** Now we heuristically extend this same tractable OU approximation with a normal steady-state distribution to more general cases. First, when  $\mu_{1,2} \neq \mu_{2,2}$ , we again act as if all agents are busy all the time. The total service rate at time  $t$  is then  $m_1\mu_{1,1} + Z_{1,2}(t)\mu_{1,2} + (m_2 - Z_{1,2}(t))\mu_{2,2}$ . To obtain the desired constant rate, we act as if  $Z_{1,2}(t)$  is constant, assuming its determined deterministic steady-state fluid approximation. This is a heuristic approximation, because we are ignoring the stochastic fluctuations in  $Z_{1,2}$ . Experiments show that this simple approximation works pretty well, but as  $n \rightarrow \infty$  in the ED regime the infinitesimal mean of the scaled queue-length process does in fact depend on the stochastic behavior of the scaled version of the stochastic process  $Z_{1,2}$  (as we would expect); i.e., simulations show that this heuristic extension is *not* asymptotically correct as  $n \rightarrow \infty$ .

We also treat the abandonments in a similar way when  $\theta_1 \neq \theta_2$ . We will approximate by a constant abandonment rate applying to all customers. For this step we also will invoke SSC (ignoring the

difference), and assume that  $Q_1(t) \approx rQ_\Sigma(t)/(1+r)$  (and similarly for  $Q_2$ ). Thus our approximating constant abandonment rate to apply to the total queue length is  $\theta \approx (r\theta_1/(1+r)) + (\theta_2/(1+r))$ . With the new approximating total service rate and average abandonment rate, we again are in the domain of an OU approximation, with normal steady-state distribution. Paralleling our previous analysis, we obtain a new approximate variance for the total queue length,

$$\text{Var}(Q_\Sigma) \approx \frac{(1+r)(\lambda_1 + \lambda_2)}{(r\theta_1 + \theta_2)}. \quad (\text{EC.3})$$

Then SSC again gives a joint normal distribution for  $(Q_1, Q_2)$  with correlation 1. The individual variances are thus approximated by (20).

## EC.2. Simulation Experiments to Evaluate the Normal Approximations

### EC.2.1. The Unbalanced-Overload Case

We now compare the approximating normal steady-state distributions to simulation results. We again consider the base case in Table 1 with  $\lambda_1 = 1.3n$  and  $\lambda_2 = 0.9n$ . The results are given in Table EC.1.

We give the standard-deviations of the total queue length  $Q_\Sigma = Q_1 + Q_2$  as well as the two queues. As before, we treat both the actual values and the scaled values, but now we are scaling in diffusion scale (dividing by  $\sqrt{n}$  after subtracting the order- $O(n)$  mean), as in (5), so that we will be substantiating the stochastic-process limit in (6). To further substantiate both the stochastic-process limit and the normal approximations, we also give the quantiles of the scaled queue lengths  $\hat{Q}_1$  and  $\hat{Q}_2$ . To save space, we omit the confidence intervals for the scaled standard deviations; these can be computed from the confidence intervals of the actual queues by dividing the half widths by  $\sqrt{n}$ .

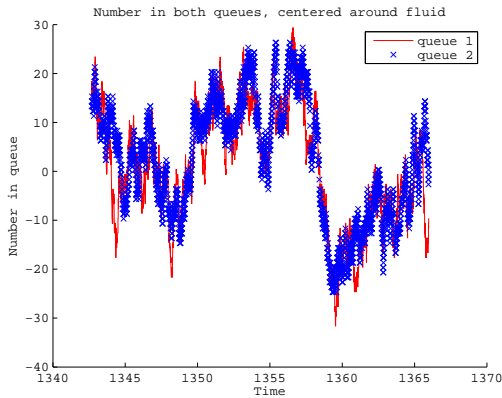
We also give the quantiles for the centered steady-state queue difference  $\tilde{D} \equiv D - E[D]$ . (Table 1 already showed that the approximation for the mean  $E[D]$  is accurate for  $n \geq 100$ .) The approximate distribution of  $D$  is obtained from the BD process in §4.1. The quantiles of the distribution of  $\tilde{D}$  pose a problem, since  $D$  is integer-valued. We thus calculate a linear interpolation of two values.

For example, for the 0.05 quantile, we took the largest value  $d_0$  such that  $P(\tilde{D} \leq d_0) < 0.05$  and linearly interpolate this value with the smallest value  $d_1$  such that  $P(\tilde{D} \leq d_1) > 0.05$ . The linear interpolation becomes just the weighted average of the two values  $d_0$  and  $d_1$ . As in Table 1,  $\tilde{D}$  is not scaled by any division.

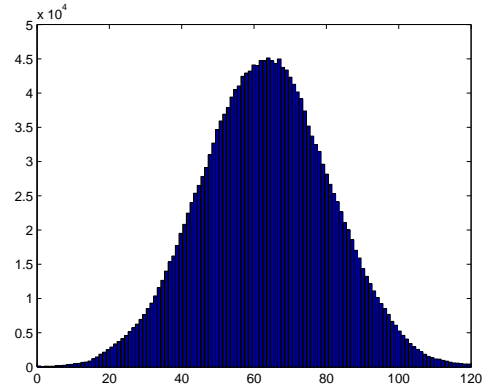
		n=25		n=100		n=400	
perf. meas.		Approx.	Sim.	Approx.	Sim.	Approx.	Sim.
$std(Q_\Sigma)$		16.6	16.0 $\pm 0.3$	33.2	33.7 $\pm 1.4$	66.3	67.6 $\pm 2.9$
$std(\hat{Q}_\Sigma)$		3.32	3.21	3.32	3.37	3.32	3.38
$std(Q_1)$		8.3	8.8 $\pm 0.1$	16.6	17.2 $\pm 0.7$	33.2	33.9 $\pm 1.4$
$std(\hat{Q}_1)$		1.66	1.75	1.66	1.72	1.66	1.7
$std(Q_2)$		8.3	8.6 $\pm 0.1$	16.6	17.1 $\pm 0.7$	33.2	33.9 $\pm 1.5$
$std(\hat{Q}_2)$		1.66	1.73	1.66	1.71	1.66	1.69
$\hat{Q}_1$ quantiles	0.05	-2.72	-2.75 $\pm 0.06$	-2.72	-2.84 $\pm 0.11$	-2.72	-2.72 $\pm 0.19$
	0.25	-1.12	-1.27 $\pm 0.08$	-1.12	-1.14 $\pm 0.03$	-1.12	-1.18 $\pm 0.08$
	0.75	1.12	1.13 $\pm 0.08$	1.12	1.14 $\pm 0.08$	1.12	1.11 $\pm 0.08$
	0.95	2.72	2.97 $\pm 0.11$	2.72	2.82 $\pm 0.20$	2.72	2.92 $\pm 0.16$
$\hat{Q}_2$ quantiles	0.05	-2.72	-2.94 $\pm 0.14$	-2.72	-2.82 $\pm 0.15$	-2.72	-2.68 $\pm 0.21$
	0.25	-1.12	-1.18 $\pm 0.08$	-1.12	-1.14 $\pm 0.04$	-1.12	-1.17 $\pm 0.06$
	0.75	1.12	1.18 $\pm 0.07$	1.12	1.14 $\pm 0.09$	1.12	1.11 $\pm 0.08$
	0.95	2.72	2.90 $\pm 0.10$	2.72	2.80 $\pm 0.20$	2.72	2.91 $\pm 0.15$
centered $D$ quantiles	0.05	-17.4	-13.4 $\pm 0.7$	-18.4	-16.6 $\pm 0.6$	-19.5	-18.2 $\pm 0.6$
	0.25	-7.4	-6.0 $\pm 0.0$	-8.4	-7.6 $\pm 0.6$	-8.5	-8.0 $\pm 0.0$
	0.75	-1.4	-0.8 $\pm 0.6$	-1.4	-1.0 $\pm 0.1$	-1.4	-1.0 $\pm 0.0$
	0.95	0.5	5.0 $\pm 1.8$	0.5	1.0 $\pm 0.1$	0.5	1.0 $\pm 0.0$

**Table EC.1** A comparison of the approximating distributions of steady-state performance measures in the unbalanced-overload case with simulation results for the base case with  $\lambda_1 = 1.3n$  and  $\lambda_2 = 0.9n$ .

To further illustrate how the approximations perform, we show two figures based on a simulation run with  $n = 100$  (the second case in Table EC.1). To show that SSC actually occurs with FQR-T, we show a plot of a segment of the queue-length sample paths in Figure EC.1. We have centered about the means, so that the average difference should be zero. To visually justify the normal approximation, we show a histogram of the class-1 queue-length distribution in Figure EC.2.



**Figure EC.1** State-space collapse for  $n = 100$ .



**Figure EC.2** Histogram of  $Q_1$  for  $n = 100$ .

### EC.2.2. The Balanced-Overload Case

In the balanced-overload case, we compare the simulation results to approximations based on the assumption that the two queues operate independently. The approximations are calculated using Whitt (2004). We simulated the same systems as before, again changing only the arrival rates to  $\lambda_1 = \lambda_2 = 1.2n$ . The results are shown in Table EC.2.

Of course, the two queues are not actually completely independent, because agents are sometimes serving customers from the other class. Thus there is some dependency between the queues. This causes the queues to be somewhat larger than if the two service pools were operating independently, because serving the other class is done somewhat inefficiently. However, the sharing is not altogether bad: although there is some increase in the queues sizes (as shown in Table 3 of the supplement), we also gain by decreasing the variance of the queues. From the efficiency point of view, we see a

tradeoff between the economies of scale provided by the sharing and the inefficiency caused by the slower service rates when sharing.

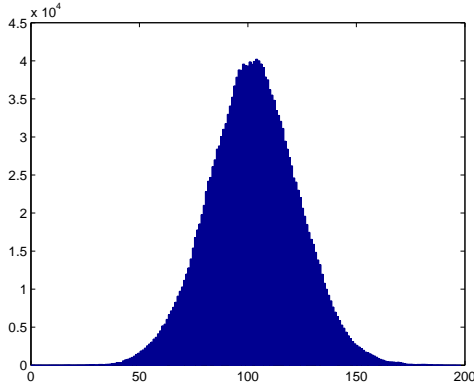
For this reason, the simulations do not match the approximations precisely, but the differences are not large. Indeed, to show the differences more clearly, for the case  $n = 100$  we added another column of simulation results for an  $M/M/100 + M$  system having a Poisson arrival process with rate  $\lambda = 120$ . Table EC.2 shows that the simulation results for this case are much closer to the approximations, showing that the errors, though not large, are not due to stochastic fluctuations. To further illustrate the difference (and the resemblance), we show histograms of the distribution of  $Q_1$  in the two cases in Figures EC.3 and EC.4. It is easy to see that the queues in both systems have a distribution that is very close to a normal distribution, and that the variance of the queue in the  $X$  model is smaller than the variance of the queue in the  $M/M/100 + M$  system.

		n=25		n=100			n=400	
perf. meas.		Approx.	Sim.	Approx.	Sim. $X$ model	Sim. Ind.	Approx.	Sim.
$std(Q_\Sigma)$		17.3	17.3 $\pm 0.5$	34.6	33.2 $\pm 2.6$	34.7 $\pm 0.7$	69.3	63.9 $\pm 3.9$
$std(\hat{Q}_\Sigma)$		3.46	3.46	3.46	3.32	3.47	3.46	3.20
$std(Q_1)$		12.2	10.37 $\pm 0.2$	24.5	20.4 $\pm 1.3$	24.7 $\pm 0.64$	49.0	38.1 $\pm 1.7$
$std(\hat{Q}_1)$		2.45	2.07	2.45	2.04	2.47	2.45	1.91
$\hat{Q}_1$ quantiles	0.05	-4.03	-3.38 $\pm 0.07$	-4.03	-3.35 $\pm 0.26$	-4.07 $\pm 0.13$	-4.03	-3.17 $\pm 0.26$
	0.25	-1.65	-1.38 $\pm 0.07$	-1.65	-1.37 $\pm 0.11$	-1.73 $\pm 0.09$	-1.65	-1.24 $\pm 0.09$
	0.75	1.65	1.30 $\pm 0.07$	1.65	1.35 $\pm 0.07$	1.69 $\pm 0.10$	1.65	1.30 $\pm 0.12$
	0.95	4.03	3.50 $\pm 0.07$	4.03	3.45 $\pm 0.20$	4.14 0.09	4.03	3.07 $\pm 0.05$

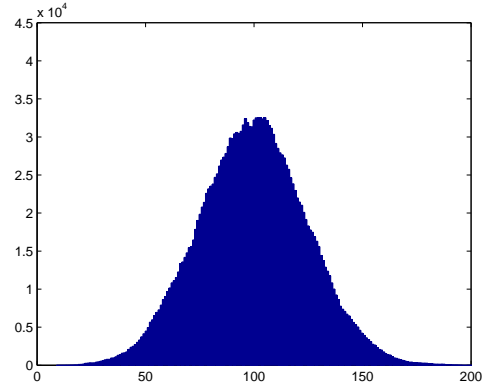
**Table EC.2** A comparison of approximations for the standard deviations and quantiles of the steady-state queue lengths with simulation estimates in the balanced-overload case with  $\lambda_1 = \lambda_2 = 1.2n$ .

### EC.3. Analytically Showing the Bad Behavior Without Thresholds

In §4.1 and Perry and Whitt (2009a) and §EC.2 of the associated e-Companion, we applied simulation to demonstrate the need for the thresholds and one-way sharing (allowing sharing in only



**Figure EC.3** A histogram for  $Q_1$  in the balanced-overload case with  $n = 100$ ,  $\lambda_i = 1.2n$  and  $\kappa_{1,2} = 10$ .



**Figure EC.4** A histogram for the queue length in an  $M/M/100 + M$  model.

one direction at a time) in FQR-T. We showed that basic FQR can perform poorly under normal loads, because it can activate sharing simultaneously in both directions that makes the system overloaded under the common assumption of inefficient sharing, as in (1). When customers do not abandon, the system becomes unstable and explodes because of the inefficient sharing. Indeed, the queue lengths tend to grow directly proportional to  $t$ . When customers do abandon, in both service pools a significant proportion of agents are serving customers of the other class. As a consequence, the queue lengths stabilize at undesirably high levels.

We now show that we can predict this performance analytically, employing a variant of the fluid model just developed. This requires that we allow two-way sharing, with both pools simultaneously serving customers of the other class. To treat FQR, we would have no thresholds at all, but the same bad behavior can occur if there are two thresholds, both of which are too small. We will thus first develop a more general six-dimensional ODE. Afterwards, we will specialize to the specific example considered in Perry and Whitt (2009a) in order to obtain closed-form expressions for key performance measures.

### EC.3.1. The Fluid-Model with Two Thresholds and Two-Way Sharing

We will now introduce the system of six deterministic fluid-model ODE's to approximate the evolution (transient behavior) of the CTMC  $X(t) \equiv (Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2)$  in the fully overloaded case considered in §3.1 when there is two-way sharing with two thresholds. As before, we focus on  $D(t) \equiv D_{1,2}(t)$ , considering the fully overloaded cases, with class 1 more overloaded than class 2. As before, the queue-difference process  $D(t)$  oscillates around the threshold  $\kappa_{1,2}$ . As before, we rely on the AP.

**The Fast-Time-Scale Process.** As before, let  $D_t(\infty)$  denote a random variable with the time-dependent steady-state distribution of the fast-time-scale process. We now exploit the time-dependent probabilities

$$\pi_{1,2}(X(t)) \equiv P(D_t(\infty) > \kappa_{1,2}) \quad \text{and} \quad \pi_{2,1}(X(t)) \equiv P(D_t(\infty) < -\kappa_{2,1}) \quad t \geq 0, \quad (\text{EC.4})$$

which will depend on the state of the approximating ODE  $X(t)$  at time  $t$ .

Again, the AP allows us to regard  $D_t$  approximately as a pure-jump continuous-time Markov process (MP), with state space  $\{k + rj : k \in \mathbb{Z}, j \in \mathbb{Z}\}$ , with transition rates that depend only on the fluid-model state at time  $t$ . There are four possible transitions in each state:  $\pm 1$  and  $\pm r$ . As before, we obtain simplification without practical sacrifice by assuming that  $r$  is rational. For rational  $r \equiv j/k$ , this is a continuous-time Markov chain (CTMC) on the state space  $\{j/k : j \in \mathbb{Z}\}$ . We then multiply by  $k$  to obtain an integer state space. Moreover, then the CTMC can be represented as a homogeneous quasi-birth-and-death (QBD) process, as in Definition 1.3.1 and §6.4 of Latouche and Ramaswami (1999). (Now the middle region  $[-\kappa_{2,1}, \kappa_{1,2}]$  is treated as part of the boundary.) For each  $t$ , we can apply the logarithmic reduction algorithm in §8.7 of Latouche and Ramaswami (1999) to efficiently calculate the steady-state distribution of  $D_t$ , i.e., the distribution of  $D_t(\infty)$ . As a consequence, we can calculate the the desired probabilities  $\pi_{i,j}(X(t))$  in (EC.4) for any fluid state  $X(t)$ .

We now specify the transition rates of the CTMC  $D_t$  given the time  $t$  and the state  $X(t)$ , using the integer state space. Let  $\lambda_+^{(j)}(m, X(t))$ ,  $\lambda_+^{(k)}(m, X(t))$ ,  $\mu_+^{(j)}(m, X(t))$  and  $\mu_+^{(k)}(m, X(t))$

be the transition rates of the FTSMC  $D_t$  for transitions of  $+j$ ,  $+k$ ,  $-j$  and  $-k$ , respectively, when  $D_t(s) = m > \kappa_{1,2}$ . Similarly, we define the transitions when  $D_t(s) = m < -\kappa_{2,1}$ :  $\lambda_-^{(j)}(m, X(t))$ ,  $\lambda_-^{(k)}(m, X(t))$ ,  $\mu_-^{(j)}(m, X(t))$  and  $\mu_-^{(k)}(m, X(t))$ , and the transitions when  $D_t(s) = m$ ,  $-\kappa_{2,1} \leq m \leq \kappa_{1,2}$ :  $\lambda_0^{(j)}(m, X(t))$ ,  $\lambda_0^{(k)}(m, X(t))$ ,  $\mu_0^{(j)}(m, X(t))$  and  $\mu_0^{(k)}(m, X(t))$ .

There are different formulas for these constant rates in the three regions. First, for  $D_t(s) = m \in [-\kappa_{2,1}, \kappa_{1,2}]$ , the upward rates are

$$\lambda_0^{(k)}(m, X(t)) = \lambda_1 \quad \text{and} \quad \lambda_0^{(j)}(m, X(t)) = \mu_{1,2}Z_{1,2}(t) + \mu_{2,2}Z_{2,2}(t) + \theta_2Q_2(t), \quad (\text{EC.5})$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 customer queue, caused by a type-2 agent service completion or by a class-2 customer abandonment. Similarly, the downward rates are

$$\mu_0^{(k)}(m, X(t)) = \mu_{1,1}Z_{1,1}(t) + \mu_{2,1}Z_{2,1}(t) + \theta_1Q_1(t) \quad \text{and} \quad \mu_0^{(j)}(m, X(t)) = \lambda_2, \quad (\text{EC.6})$$

corresponding, first, to a departure from the class-1 customer queue, caused by a class-1 agent service completion or by a class-1 customer abandonment, and, second, to a class-2 arrival.

Next, for  $D_t(s) = m \in (\kappa_{1,2}, \infty)$ , we have upward rates

$$\lambda_+^{(k)}(m, X(t)) = \lambda_1 \quad \text{and} \quad \lambda_+^{(j)}(m, X(t)) = \theta_2Q_2(t), \quad (\text{EC.7})$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 customer queue caused by a class-2 customer abandonment. The downward rates in this region are

$$\mu_+^{(k)}(m, X(t)) = \mu_{1,1}Z_{1,1}(t) + \mu_{2,1}Z_{2,1}(t) + \mu_{2,2}Z_{2,2}(t) + \theta_1Q_1(t) \quad \text{and} \quad \mu_+^{(j)}(m, X(t)) = \lambda_2, \quad (\text{EC.8})$$

corresponding, first, to a departure from the class-1 customer queue, caused by (i) a type-1 agent service completion, (ii) a type-2 agent service completion, or (iii) by a class-1 customer abandonment and, second, to a class-2 arrival.

Finally, for  $D_t(s) = m \in (-\infty, -\kappa_{2,1})$ , the upward rates are  $\lambda_-^{(k)}(m, X(t)) \equiv \lambda_1$  and

$$\lambda_-^{(j)}(m, X(t)) \equiv \mu_{2,2}Z_{2,2}(t) + \mu_{1,2}Z_{1,2}(t) + \mu_{2,1}Z_{2,1}(t) + \mu_{1,1}Z_{1,1}(t) + \theta_2Q_2(t), \quad (\text{EC.9})$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 customer queue, caused by (i) a type-1 agent service completion, (ii) a type-2 agent service completion, or (iii) by a class-2 customer abandonment. The downward rates in this region are

$$\mu_-^{(k)}(m, X(t)) \equiv \theta_1 Q_1(t) \quad \text{and} \quad \mu_-^{(j)}(m, X(t)) \equiv \lambda_2, \quad (\text{EC.10})$$

corresponding, first, to a departure from the class-1 customer queue, caused by a class-1 customer abandonment, and, second, a class-2 arrival.

For each time  $t$  and state  $X(t)$ , we can apply the QBD algorithm with the rates above to find the distribution of  $D_t(\infty)$  and then the important quantities  $\pi_{1,2}(X(t))$  and  $\pi_{2,1}(X(t))$  in (EC.4).

**The ODE's.** As in §4, we are considering the fully-overloaded case in §3.1, so that the arrival rates are sufficiently high that both approximate queue lengths are positive in steady state. We again consider the transient behavior of the fluid model under the assumption that all agents are busy. We are thus describing the transient behavior in the second transient period, leading to steady state.

First, given  $Z_{i,j}(t)$  and  $\pi_{i,j}(X(t))$ , we obtain ODE's for the two queue-length processes. As before, we let the derivative  $\dot{Q}_1(t)$  equal the rate of increase of  $Q_1(t)$  minus its rate of decrease. The rate of increase is simply the arrival rate to customer queue 1,  $\lambda_1$ . The rate of decrease is more complicated. First, there is the rate of abandonment from queue 1, which is  $Q_1(t)\theta_1$ . Second, there is the rate of decrease from queue 1 due to service completions by servers who will next take customers from queue 1, which depends on the state of the queue-difference stochastic process. Exploiting the AP, we will not focus on the actual state of the queue-difference process, but instead focus on the average state, assuming that the queue-difference process oscillates relatively rapidly compared to the other processes. We thus assume that a proportion  $\pi_{1,2}(X(t))$  of the time the threshold  $\kappa_{1,2}$  is exceeded. That portion of the decrease rate is  $\pi_{1,2}(X(t)) (Z_{1,1}(t)\mu_{1,1} + Z_{2,1}(t)\mu_{2,1} + Z_{1,2}(t)\mu_{1,2} + Z_{2,2}(t)\mu_{2,2})$ . There will be corresponding, but different, rates of decrease for the proportion of time  $\pi_{2,1}(X(t))$  that the queue-difference is below the lower threshold  $-\kappa_{2,1}$ , and the proportion of time  $1 - \pi_{1,2}(X(t)) - \pi_{2,1}(X(t))$  that the queue-difference is between the two thresholds. Note that  $Q_1$  is

decreased due to service completions only when  $D_t > -\kappa_{2,1}$ , i.e., a proportion of  $1 - \pi_{2,1}(X(t))$  of the time. Similarly,  $Q_2$  is decreased due to service completions only when  $D_t < \kappa_{1,2}$ , or a proportion  $1 - \pi_{1,2}(X(t))$  of the time. That reasoning leads to the two ODE's for the queue-length processes:

$$\begin{aligned}\dot{Q}_1(t) &= \lambda_1 - Q_1(t)\theta_1 - \pi_{1,2}(X(t)) (Z_{1,1}(t)\mu_{1,1} + Z_{2,1}(t)\mu_{2,1} + Z_{1,2}(t)\mu_{1,2} + Z_{2,2}(t)\mu_{2,2}) \\ &\quad - (1 - \pi_{1,2}(X(t)) - \pi_{2,1}(X(t)) (Z_{1,1}(t)\mu_{1,1} + Z_{2,1}(t)\mu_{2,1})) \\ \dot{Q}_2(t) &= \lambda_2 - Q_2(t)\theta_2 - \pi_{2,1}(X(t)) (Z_{2,2}(t)\mu_{2,2} + Z_{1,2}(t)\mu_{1,2} + Z_{2,1}(t)\mu_{2,1} + Z_{1,1}(t)\mu_{1,1}) \\ &\quad - (1 - \pi_{2,1}(X(t)) - \pi_{1,2}(X(t)) (Z_{2,2}(t)\mu_{2,2} + Z_{1,2}(t)\mu_{1,2})) .\end{aligned}\tag{EC.11}$$

Assuming that the system is indeed fully overloaded, all the servers will be working, so that

$$Z_{1,1}(t) + Z_{2,1}(t) = m_1 \quad \text{and} \quad Z_{1,2}(t) + Z_{2,2}(t) = m_2 .\tag{EC.12}$$

Now we propose approximating ODE's for  $Z_{i,j}(t)$ , exploiting the proportions  $\pi_{i,j}(X(t))$ :

$$\begin{aligned}\dot{Z}_{1,1}(t) &= (1 - \pi_{2,1}(X(t))Z_{2,1}(t)\mu_{2,1} - \pi_{2,1}(X(t))Z_{1,1}(t)\mu_{1,1}) \\ \dot{Z}_{1,2}(t) &= \pi_{1,2}(X(t))Z_{2,2}(t)\mu_{2,2} - (1 - \pi_{1,2}(X(t))Z_{1,2}(t)\mu_{1,2})\end{aligned}\tag{EC.13}$$

Combining §4.1, equations (EC.11)–(EC.13) and appropriate initial conditions, we have a specification of the transient dynamics of the fluid model, assuming that the system remains fully overloaded: a system of six ODE's or one six-dimensional ODE.

**Steady state.** We now characterize the steady state of this six-dimensional fluid model in the fully overloaded case as the solution to eight equations. For that purpose, let  $Q_i \equiv Q_i(\infty)$ ,  $Z_{i,j} \equiv Z_{i,j}(\infty)$  and let  $D \equiv D_\infty$  be the queue-difference process when the system is in steady state (for simplicity, we use the same notation). Hence,  $D(\infty)$  stands for the steady-state distribution of  $D_\infty$ , i.e., the steady-state of the weighted-difference process when the system itself is in steady state. As before, we define

$$\pi_{1,2} \equiv P(D(\infty) > \kappa_{1,2}) \quad \text{and} \quad \pi_{2,1} \equiv P(D(\infty) < -\kappa_{2,1}).\tag{EC.14}$$

We can find the steady-state values  $Q_i$  and  $Z_{i,j}$ ,  $i, j = 1, 2$ , by simply setting the derivatives on the left sides of the ODE's in (EC.11) and (EC.13) equal to 0. We then have to solve the 8 equations (EC.11) – (EC.14) with the 8 unknowns: the two  $Q_i$ , four  $Z_{i,j}$  and the two  $\pi_{i,j}$ .

Solving these eight equations is not easy since two of them, (EC.14), involve solving the MP balance equations or applying a variant of the QBD algorithm in Perry and Whitt (2009b). To solve these eight equations, we propose an **iterative algorithm**: We start by setting  $Z_{1,1} = m_1$ ,  $Z_{2,2} = m_2$  and  $Q_1 = Q_2 = 0$ . We then solve numerically the balance equations to find  $\pi_{1,2}$  and  $\pi_{2,1}$ . We can then use these two  $\pi_{i,j}$  in the other  $Q_i$  and  $Z_{i,j}$  equations to get their new values. We then use these values of  $Q_i$  and  $Z_{i,j}$  in the balance equations to solve for the  $\pi_{i,j}$  again, and keep iterating until convergence. Extensive numerical experience indicates that this algorithm converges rapidly to values that are extremely close to simulation results. (It typically takes only a few iterations.)

### EC.3.2. A Symmetric Special Case

Having developed the more general approximate fluid model for the overloaded X model with the FQR-T control, allowing two-way sharing and having two thresholds, we now want to consider the specific case of pure FQR as in Gurvich and Whitt (2009a,b,c) and §4.1 of Perry and Whitt (2009a). As pointed out in Perry and Whitt (2009a), the bad behavior should not be unexpected, because the  $X$  model with service rates depending on *both* the customer class and the service pool violates the conditions in the theorems in Gurvich and Whitt (2009a,b,c).

In what follows, we simplify the equations by considering a **symmetric system**. That will allow us to obtain explicit expressions for the performance measures.

In the symmetric model we have  $r = 1$ . Then the MP  $\{D_t(s) : s \geq 0\}$  becomes a birth-and-death (BD) process. Then FQR reduces to the policy of *servicing the longer queue* (SLQ). It also reduces to our FQR-T control with thresholds set at  $\kappa_{1,2} = \kappa_{2,1} = 0$  for all  $n$ , without imposing the constraint of one-way sharing. (That still leave three regions:  $(-\infty, 0)$ ,  $\{0\}$  and  $(0, \infty)$ ). Throughout this section we will refer to this policy as the SLQ policy.

For the SLQ policy, it is important to specify how ties are broken, because the control tends to make ties occur quite frequently. Here we are assuming that a server will always serve a customer from his own class if the queue lengths are equal. The symmetry implies that  $\pi_{1,2} = \pi_{2,1}$ ,  $Z_{1,2} = Z_{2,1}$  and  $Q_1 = Q_2$ .

We first consider the case of no customer abandonment, and then afterwards the case of customer abandonment. For both, we will give numerical results for a symmetric  $X$  model with parameters:  $m_i = n$ ,  $\lambda_i = 0.99n$ ,  $\mu \equiv \mu_{i,i} = 1$  and  $\nu \equiv \mu_{1,2} = \mu_{2,1} = 0.8$ , where inefficiency occurs because  $\nu < \mu$  (serving the other class is less efficient). With these parameters, if each service pool served only its own class, then the system would be stable, even without abandonments.

**Without abandonment.** We now show how to analyze such a symmetric  $X$  model with the SLQ routing policy. To do so, we will work in the fluid scaling, dividing by  $n$ . For that purpose, let  $z(t) \equiv Z_{i,i}(t)/n$ , be the *proportion* of agents serving their own class in each of the pools, and let  $q(t) \equiv Q_i(t)/n$ . Since we consider the systems normalized by  $n$ , we take  $\lambda \equiv \lambda_i/n$ , so that in our example above  $\lambda = 0.99$ . Because of the symmetry, we omit the class subscripts. To preserve the symmetry, we assume that the initial conditions are symmetric as well.

We now develop ODE's describing the evolution of  $z(t)$  and  $q(t)$ . The reasoning is very similar to before, but we will obtain significant simplification by exploiting the symmetry. We first find the time-dependent proportion of time that the two queues are equal. Let  $\pi(t)$  be that proportion, i.e.,  $\pi(t) = P(D_t(\infty) = 0)$ ,  $t \geq 0$ . (See §4.1 for more details.) By symmetry, the amount of time queue 1 is bigger than queue 2 is equal to the amount of time queue 1 is smaller than queue 2. Hence,  $(1 - \pi(t))/2$  is the amount of time queue 1 is bigger than queue 2.

We first write down the ODE for  $z$ . It is easy to see that

$$\dot{z}(t) = \pi(t)(1 - z(t))\nu + \frac{1 - \pi(t)}{2} [\nu - z(t)(\mu + \nu)]. \quad (\text{EC.15})$$

In equilibrium,  $\dot{z}(\infty) = 0$ , so that we get

$$z \equiv z(\infty) = \frac{\nu(1 + \pi)}{2\nu\pi + (1 - \pi)(\mu + \nu)}, \quad (\text{EC.16})$$

where  $\pi \equiv \pi(\infty)$ . In our numerical example with  $\mu = 1$  and  $\nu = 0.8$ , equation (EC.16) becomes  $z = (4 + 4\pi)/(9 - \pi)$ .

To find the value of  $z$  above, we need to solve for  $\pi$ . We will find an expression for  $\pi$  by approximating the *absolute* difference process between the queues:  $\{|D(t)| : t \geq 0\}$  by a BD process. For all states  $j \geq 1$ , the birth rate is  $\hat{\lambda}_j = \lambda$ , corresponding to an arrival at the larger queue, while the death rate is  $\hat{\mu}_j = \lambda + 2[z\mu + (1 - z)\nu]$ , corresponding to an arrival to the shorter queue, or any service completion (since the newly available agent will take a customer from the larger queue). There is a different birth rate when the two queues are equal. The birth rate (to make either of the queues longer) when the difference is zero is  $\hat{\lambda}_0 = 2\lambda + 2[z\mu + (1 - z)\nu]$ , where  $2\lambda$  corresponds to an arrival to either of the queues, while  $2[z\mu + (1 - z)\nu]$  corresponds to any service completion. Solving for the steady-state of this BD process, we get  $\pi = (1 - \rho)/(1 - \rho + (\hat{\lambda}_0/\hat{\mu}))$ , where

$$\rho \equiv \frac{\hat{\lambda}_j}{\hat{\mu}_j} = \frac{\lambda}{\lambda + 2[z\mu + (1 - z)\nu]}, \quad j \neq 0. \quad (\text{EC.17})$$

Hence we get

$$\pi = \frac{z\mu + (1 - z)\nu}{\lambda + 2z\mu + 2(1 - z)\nu}. \quad (\text{EC.18})$$

Solving the two equations (EC.16) and (EC.18) for  $\pi$  and  $z$  with the rates of our numerical example, we get  $z = 0.61$  ( $Z_{i,i} = 61$ ) and  $\pi = 0.32$ .

Using the values of  $z$  just determined, we can now find the deterministic fluid approximation for the evolution of the queues, after  $z$  achieves its steady-state, and is fixed. In the fluid approximation work flows to and out of the system in constant rate, hence the rate of change in the queue length is the arrival rate minus the departure rate, yielding:  $\dot{q}(t) = \lambda - z\mu - (1 - z)\nu$ , so that

$$q(t) = q(0) + [\lambda - \nu - (\mu - \nu)z]t, \quad t \geq 0.$$

Applying the rates from our numerical example, we get  $q(t) = q(0) + 0.068t$  or  $Q(t) = Q(0) + 6.8t$ ,  $t \geq 0$ .

To evaluate the accuracy of the approximations above, we performed simulation experiments. As before, we used 5 independent runs, each with 300,000 arrivals. We report averages together with

the half widths of the 95% confidence intervals, based on a  $t$  statistic with four degrees of freedom. Here the simulation results are: slope of  $E[Q_1(t)] = 6.8 \pm 0.4$ ,  $E[Z_{1,1}] = 61.2 \pm 0.6$ ,  $\pi = 0.33 \pm 0.00$ , while the respective approximations are 6.8, 61.0 and 0.32. Figures in §4.1 of Perry and Whitt (2009a) compare the sample paths of  $Q_1(t)$  and  $Z_{2,1}(t)$ , starting empty, in one simulation run. The queue-length sample path and the straight line are almost indistinguishable.

**With Abandonment.** When we consider the symmetric model with abandonments, we have a new ODE for the queue length:

$$\dot{q}(t) = \lambda - z(t)\mu - (1 - z(t))\nu - q(t)\theta,$$

where  $\theta \equiv \theta_i$ ,  $i = 1, 2$  is the abandonment rate for both classes, which we take to be 0.2 in our example. In steady-state, we have  $\dot{q}(t) = 0$ , and thus, for  $q \equiv q(\infty)$ ,

$$q = \frac{\lambda - z\mu - (1 - z)\nu}{\theta}. \quad (\text{EC.19})$$

An initial approximation for  $q$  can use the same value of  $z$  we obtained before without abandonments. If we use  $z = 0.61$  from before (without abandonment), then we get  $q = 0.34$ , which turns out to be quite accurate. However, proceeding more carefully, we can incorporate the abandonment and investigate how it changes the value of  $z$ . We see that it does so through the way it changes the value of  $\pi$ . To do the analysis, we need to consider once again the birth and death rates of the process  $|D(\infty)|$ : Let  $\hat{\lambda}_j$  be the birth rate, and  $\hat{\mu}_j$  be the death rate when the difference between the two queues is  $j$ . We then have  $\hat{\lambda}_0 = 2\lambda + 2[z\mu + (1 - z)\nu] + 2q\theta$ ; for  $j > 0$ , the birth rates are  $\hat{\lambda}_j = \lambda + q\theta$ , while the death rates are  $\hat{\mu}_j = \lambda + 2[z\mu + (1 - z)\nu] + q\theta$ .

Solving the BD balance equations, we get an expression for  $\pi$  in terms of  $z$  and  $q$ . Once again we get  $\pi = (1 - \rho)/(1 - \rho + (\hat{\lambda}_0/\hat{\mu}))$ , but where  $\rho$  is redefined as  $\rho \equiv (\lambda + q\theta)/(\lambda + 2[z\mu + (1 - z)\nu] + q\theta)$ .

Hence

$$\pi = \frac{z\mu + (1 - z)\nu}{\lambda + 2z\mu + 2(1 - z)\nu + q\theta}.$$

From equation (EC.19), we get  $q\theta = \lambda - z\mu - (1 - z)\nu$ , thus

$$\pi = \frac{z\mu + (1 - z)\nu}{2\lambda + z\mu + (1 - z)\nu}. \quad (\text{EC.20})$$

Solving the two equations (EC.16) and (EC.20) in the two unknowns  $\pi$  and  $z$ , we get  $z = 0.607$  and  $\pi = 0.318$  in our numerical example. Inserting that value of  $z$  in equation (EC.19), we get  $q = 0.343$ , or  $Q_i = 34.3$  and  $Z_{i,i} = 60.7$ . On average, 39.3 agents are serving customers from the other class, hence the total service rate of each class reduces from  $100\mu = 100$  to  $60.7\mu + 39.3\nu = 92.14$ , which is less than the arrival rate  $\lambda_i = 99$ . That explains why the system becomes congested. The corresponding simulation estimates are:  $E[Q_1] = 34.3 \pm 0.8$ ,  $E[Z_{1,1}] = 61.0 \pm 0.0$  and  $\pi = 0.34 \pm 0.01$ . Figures showing typical sample paths are given in §3 of the online supplement.